N° d'ordre : 3467

UNIVERSITÉ DE RENNES 1

THÈSE

présentée devant

l'université de Rennes 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention INFORMATIQUE

par

Fabienne MOREAU

Équipe universitaire : TEXMEX - IRISA
École doctorale : MATISSE
Composante universitaire : IFSIC

Titre de la thèse :

Revisiter le couplage
traitement automatique des langues
et
recherche d'information

soutenue le 07 décembre 2006 devant la commission d'examen

Présidente :	Marie-Odile	CORDIER	Professeur, université de Rennes 1
Rapporteurs :	Mohand	BOUGHANEM	Professeur, université de Toulouse 3
	Jian-Yun	NIE	Professeur, université de Montréal
Examinatrice :	Adeline	NAZARENKO	Professeur, université de Paris 13
Directrice :	Pascale	SÉBILLOT	Professeur, INSA de Rennes

Remerciements

En préambule de ce mémoire, je tiens tout d'abord à exprimer mes sincères remerciements aux membres du jury qui m'ont fait l'honneur d'accepter d'évaluer mon travail de thèse.

Merci à Mme Marie-Odile CORDIER, Professeur à l'université de Rennes 1, d'avoir accepté de présider le jury de cette thèse. Merci également à M. Mohand BOUGHANEM, Professeur à l'université de Toulouse III, et à M. Jian-Yun NIE, Professeur à l'université de Montréal, pour leur difficile travail de rapporteur. Qu'ils trouvent ici le témoignage de ma reconnaissance. Merci enfin à Mme Adeline NAZARENKO, Professeur à l'université de Paris 13, pour l'intérêt qu'elle a porté à mes travaux en examinant ce mémoire.

Ces travaux et ce manuscrit n'auraient pu voir le jour sans le soutien crucial de trois personnes en particulier. Je tiens tout particulièrement à remercier Mme Pascale SÉBILLOT qui a dirigé cette thèse, pour sa disponibilité, sa patience, et ses précieux conseils qui ont été importants tout au long de cette recherche. Ses nombreuses relectures plus que pertinentes de mon manuscrit, son recul sur le sujet et sa rigueur m'ont beaucoup appris. Je remercie également Vincent CLAVEAU, dont l'aide sur le plan technique et scientifique a été plus que déterminante pour mener ce travail à terme. Je le remercie également pour son écoute attentive et ses qualités humaines. Je souhaite enfin remercier Laurent AMSALEG pour l'intérêt qu'il a su porter à ce travail, pour le temps consacré à réfléchir avec nous et à poser les bonnes questions.

Merci donc à ces trois chercheurs avec qui j'ai eu la chance de pouvoir travailler. J'ai beaucoup appris à leur côtés et je leur adresse toute ma reconnaissance.

Merci également à Patrick GROS, directeur de l'équipe TexMex, de m'avoir accueilli au sein de son équipe et de m'avoir fait confiance durant ces trois années.

La réalisation de ce travail s'appuie également sur un environnement qui est essentiel. À ce titre, je voudrais remercier l'INRIA et l'IRISA pour les conditions exceptionnelles dans lesquelles elles permettent aux doctorants d'évoluer. Merci également au personnel de la cafétéria (Pierrette, Laurence et Cie) pour leur bonne humeur quotidienne et leur accueil chaleureux.

Je tiens à remercier sincèrement tous les membres de l'équipe TexMex qui, chacun à leur façon, m'ont témoigné leur soutien au cours de ces trois années de cette thèse.

Merci à Cédric pour ses nombreux conseils « techniques », son aide plus que précieuse pour accélérer mes temps de traitements (!) et plus généralement pour sa gentillesse et sa grande disponibilité. Merci à Nicolas pour ses cours personnalisés de statistiques, ses conseils et ses critiques constructives. Merci également à lui d'avoir partagé avec moi les états d'âmes d'un doctorant en phase de rédaction. Merci à Zied, Stephane, Pierre-Hugues, Romain, Loïc et tous les stagiaires de passage qui par leurs encouragements ont contribué à l'aboutissement de cette thèse.

Un grand merci à mon co-bureau — mais néanmoins ami — qui se reconnaîtra, pour son soutien quotidien, ses délicates attentions, ses longues discussions (tout seul ou avec moi ;-) et la bonne ambiance qu'il a su apporter à notre bureau...

Un remerciement « spécial » à Gaëtan et à Ar'ch'ie qui, sans le savoir (!), ont fortement contribué à la recharge des batteries.

Enfin, je terminerai cette énumération (non exhaustive) en m'adressant plus particulièrement à Vincent qui a su faire preuve d'une patience sans limite (vraiment aucune !) pour me (sup-)porter dans les bons comme dans les mauvais moments. Je le remercie sincèrement ici de son aide, de son dévouement et de ses sacrifices et lui dédie ce travail.

Table des matières

Introduction

La recherche d'information (RI) est un vaste domaine d'étude apparu dans les années 60 qui a subi de constantes évolutions. Alors qu'historiquement les travaux réalisés en son sein étaient destinés à étudier et concevoir des outils de recherche réservés à une communauté de spécialistes — les premiers systèmes ont été construits afin d'aider les bibliothécaires à retrouver des documents contenus dans des bases bibliographiques —, l'avènement d'Internet et plus particulièrement du *Web* a conduit à révéler la RI au grand jour, notamment par le biais des moteurs de recherche. La profusion de données numériques disponibles a rendu indispensables des moyens de recherche performants et automatiques, permettant à tout un chacun de trouver une information précise. De la recherche documentaire proprement dite, la RI a alors évolué vers des tâches de plus en plus nombreuses et diversifiées. Les systèmes de recherche d'information (SRI) doivent aujourd'hui savoir traiter des volumes gigantesques de données, s'adapter aux nouveaux modes de communication, gérer la nature multimédia de l'information (l'image, le son, la vidéo, le texte...). Les différentes pistes de recherche offertes chaque année lors de la conférence TREC (*Text Retrieval Conference*), qui représente l'une des plus importantes rencontres scientifiques autour de l'évaluation des systèmes de recherche d'information, témoignent de cette évolution. Outre selon les axes de recherches traditionnels, la dernière campagne de 2006 proposait ainsi d'évaluer les SRI à travers la recherche dans les *blogs*, la recherche automatique de *spams* (*spam track*), la recherche spécialisée dans des données d'entreprise (*enterprise track*), dans des données juridiques (*legal track*), la recherche dans la vidéo (*video track*)... Tous ces types de systèmes de RI, quels que soient leurs objectifs respectifs, la nature ou la provenance de l'information manipulée, tendent en fait vers le même but : établir une correspondance entre l'information disponible et celle recherchée par l'utilisateur. Toute la difficulté de cette tâche de RI réside essentiellement autour de la pertinence du lien qui sera établi. Dans le cadre de ce mémoire, nous nous intéressons plus particulièrement à l'information de nature textuelle.

En RI textuelle et automatique, les méthodes traditionnellement utilisées pour faire la relation entre l'information recherchée par l'utilisateur, généralement exprimée par une requête formulée à l'aide de mots-clés ou d'une phrase en langage naturel, et l'information disponible, représentée par un ensemble de documents textuels, reposent sur une mise en correspondance des mots utilisés dans la requête avec ceux représentant le contenu des documents. Nous reviendrons plus en détail sur les fondements de la RI textuelle et les techniques disponibles dans le chapitre 1 de ce mémoire. La pertinence

de l'appariement entre une requête et un document repose donc essentiellement sur une comparaison de mots. Compte tenu de ce mécanisme, les SRI se retrouvent rapidement confrontés à deux problèmes importants. Le premier est lié à la possibilité offerte par le langage naturel de formuler de différentes manières un même concept, une même idée. Un document pertinent peut contenir des termes[1] « sémantiquement » proches de ceux de la requête mais toutefois différents (*e.g.* des synonymes). Ainsi, un document composé par exemple du terme *automobile* ne pourra être apparié à une requête représentée par le mot *voiture*. Ce phénomène provoque une baisse des performances des systèmes qui ne peuvent pas proposer à un utilisateur des documents pourtant intéressants. Le second problème, dual du premier, est lié au caractère fortement polysémique des mots de la langue : un même terme peut renvoyer à des concepts totalement différents. L'ambiguïté des mots qui en découle entraîne alors potentiellement la récupération de documents non pertinents par le SRI. Un utilisateur qui recherche par exemple des informations sur le métier d'*avocat* (la profession juridique) peut ainsi se voir retourner des documents qui concernent le fruit. Le mécanisme d'appariement tel qu'il est proposé par les méthodes classiques de RI trouve donc ses limites dans la complexité du langage naturel.

L'une des solutions souvent envisagée pour faire face à ces difficultés est d'intégrer au sein des SRI une analyse linguistique des documents et des requêtes, qui présente l'avantage de ne plus considérer les mots comme de simples graphies mais comme des entités linguistiques à part entière, *i.e.* des unités susceptibles d'avoir plusieurs sens, de subir des variations de forme (*e.g.* les formes *cheval* et *chevaux*), de structure (*e.g.* les expressions *appartement à vendre* et *vente d'appartement*)... Une telle analyse est effectuée en RI par le biais de techniques du traitement automatique des langues (TAL). Le rôle des traitements linguistiques en RI est d'extraire automatiquement des informations linguistiques des documents et requêtes qui doivent permettre aux SRI une meilleure préhension des contenus textuels et, par conséquent, les aider à retrouver de manière plus pertinente ce que recherche l'utilisateur.

Le couplage de la RI et de techniques du TAL (par le biais des informations qu'elles permettent d'extraire) apparaît donc comme une solution assez naturelle pour améliorer la pertinence de l'appariement requête-documents. De nombreuses études se sont intéressées à cette association. Le deuxième chapitre de ce mémoire propose, à travers un état de l'art détaillé de ces différents travaux, une synthèse des principales contributions du TAL à la RI. À la suite de cette analyse, il reste cependant difficile de déterminer l'apport réel de l'exploitation d'informations linguistiques en RI. En effet, comme nous le montrons, les résultats obtenus dans les diverses expérimentations de couplage déjà proposées sont souvent contradictoires. Certaines expériences mettent en évidence l'intérêt de prendre en compte des informations linguistiques en RI, constatant une amélioration significative des performances des SRI par rapport à des systèmes traditionnels. D'autres au contraire conduisent à des résultats plus mitigés. Pour beaucoup, l'impact des informations linguistiques est très irrégulier et tributaire de nombreux paramètres. Enfin, certaines études vont même jusqu'à observer une dégradation des performances

[1] Dans ce mémoire, le mot « terme » est employé, sauf mention explicite du contraire, dans le sens de « mot », sans référence particulière à des notions de terminologie.

des SRI. Malgré le nombre et la diversité des travaux déjà réalisés, le rôle du TAL en RI demeure par conséquent toujours assez flou. Spärck-Jones, dont les travaux font référence dans ces deux domaines de recherche, se montre même plutôt pessimiste (Spärck Jones, 1999) : « *It is not clear, [either] that NLP*[2] *is required for some tasks that are closely related to ordinary retrieval.* ».

Le principal objectif des travaux présentés ici est de chercher à comprendre pourquoi la convergence entre ces deux domaines ne donne pas des résultats plus catégoriques et de se poser par conséquent de nouvelles questions quant aux méthodes à adopter pour enrichir la RI à l'aide de techniques issues du TAL. Pour cela, nous ne cherchons pas à expérimenter l'apport de nouvelles informations linguistiques en RI mais plutôt, en nous appuyant sur les résultats des travaux existants, à étudier le couplage TAL-RI sous un nouvel angle. Les hypothèses proposées ne sont donc pas nécessairement nouvelles, mais présentent la particularité d'être abordées de façon originale.

Le point de départ de nos recherches s'appuie sur le constat que la majorité des travaux qui proposent d'enrichir la RI à l'aide de techniques du TAL tente d'évaluer l'apport d'un seul type d'information linguistique. Cette information appartient généralement à l'un des trois niveaux de la langue suivants : le niveau morphologique — qui s'intéresse à la forme des mots —, syntaxique — qui concerne la façon dont les mots s'articulent entre eux pour former des syntagmes ou des phrases — ou sémantique — qui étudie le sens des mots. Une des limites de cette approche est qu'elle n'exploite que partiellement la richesse de la langue, puisqu'une description linguistique fine des textes fait nécessairement intervenir tous les niveaux de la langue. La définition de l'analyse sémantique proposée par Delafosse (1999) illustre ainsi bien l'idée d'un lien étroit entre ces trois niveaux : « l'analyse sémantique prend comme unité d'analyse la phrase et conduit à représenter sa partie significative. Ces phrases, dont l'analyseur sémantique doit décrire le sens, se composent d'un certain nombre de *mots* identifiés par l'analyse morphologique et regroupés en *structures* par l'analyse syntaxique. Ces mots et ces structures constituent autant d'indices pour le calcul du sens : on pourrait donc dire que le sens résulte de la double donnée du sens des mots et du sens des relations entre ces mots ».

Compte tenu de ce constat, et dans le but d'aborder autrement le couplage TAL-RI, nous proposons d'étudier l'impact en RI d'une analyse linguistique riche des contenus textuels des documents et requêtes qui prend en compte de manière simultanée ces trois niveaux de langue. Pour cette étude, nous ne pouvons pas cependant nous contenter de multiplier les informations linguistiques multi-niveaux au sein d'un SRI et d'évaluer leur apport sur ses performances. Nous devons au préalable nous poser un certain nombre de questions fondamentales et offrir des méthodes efficaces pour y répondre. Étant donné la nature différente des informations linguistiques prises en compte, celles-ci n'ont vraisemblablement pas toutes le même impact sur les performances des systèmes. Il est donc nécessaire, dans un premier temps, de s'interroger sur leur efficacité respective pour retrouver des documents pertinents ; cette première étude permet d'avoir une idée plus précise des résultats que l'on peut attendre de leur

[2] *Natural Language Processing*.

combinaison. Il est également important d'examiner la façon dont elles se comportent les unes par rapport aux autres. Certaines informations peuvent en effet avoir un impact similaire sur les performances (*e.g.* retrouver les mêmes documents pertinents). Dans ce cas, il s'avère inutile de multiplier des connaissances linguistiques dont l'apport en RI est identique. D'autres, au contraire, peuvent intervenir de manière complémentaire, leur couplage permettant alors de retrouver plus de documents pertinents que si ces informations étaient prises en compte individuellement. Le troisième chapitre de ce mémoire apporte des éléments de réponse à ces questions très peu abordées dans la littérature en proposant des techniques inédites pour ce faire. Outre l'état de l'art sur les principaux apports du TAL à la RI du chapitre 2, il constitue notre première contribution. Il vise plus précisément deux objectifs. Le premier est de mesurer, dans un cadre homogène et sur des données identiques, l'efficacité de chacun des différents types d'informations linguistiques pour retrouver des documents pertinents. Le second est d'obtenir une réponse quant à l'intérêt ou non d'intégrer au sein d'un même SRI de manière conjointe des informations linguistiques de plusieurs niveaux de langue. Pour ce faire, nous proposons, à partir d'une plate-forme réalisée pour intégrer au sein d'un SRI des connaissances multi-niveaux, une analyse originale des corrélations entre ces diverses informations du point de vue de leur efficacité.

Ces études révèlent des résultats intéressants et tranchés qui attestent de l'intérêt de la prise en compte d'informations linguistiques en RI et de celui du couplage de connaissances multi-niveaux. À partir de ces conclusions, il convient alors de s'interroger sur la façon de combiner ces informations au sein d'un SRI de manière à optimiser leur exploitation. Pour cela, nous proposons de fusionner les résultats obtenus par chacune des connaissances exploitées de manière individuelle par le SRI. La fusion de ces données est cependant problématique. Nos expériences précédentes montrent en effet l'hétérogénéité de l'impact des types de connaissances. On ne peut donc pas leur accorder une importance identique. Cependant, malgré cette diversité d'efficacité, on ne souhaite pas non plus fixer *a priori* l'importance à donner à chacun d'entre eux. En effet, certaines informations considérées comme globalement moins pertinentes peuvent parfois être utiles pour récupérer des documents intéressants. La méthode de fusion doit s'adapter à ces différents cas de figure. Enfin, une de nos hypothèses est que l'efficacité des diverses connaissances exploitées par le SRI est tributaire du type de requête pris en compte et, plus précisément, de la nature même des informations qui la compose. Cette hypothèse rejoint les travaux qui s'intéressent à la prédiction de la difficulté des requêtes et qui montrent l'existence de corrélations entre les informations contenues dans les requêtes et les performances des systèmes. Nous souhaitons donc également adapter l'importance à donner aux diverses informations linguistiques en fonction des spécificités des requêtes. Pour toutes ces raisons, nous ne pouvons nous appuyer sur des méthodes de fusion de données traditionnelles. Nous proposons donc une approche nouvelle, basée sur un système d'apprentissage artificiel qui, à partir des différentes listes de résultats dont nous disposons et en tenant compte des particularités des requêtes traitées, détecte la meilleure façon de fusionner ces résultats. L'application de notre méthode de fusion permet alors d'attester l'intérêt de l'exploitation conjointe de diverses

connaissances en RI. Ce système fait l'objet de notre quatrième chapitre et représente notre deuxième contribution.

Enfin, en étudiant l'intérêt du couplage d'informations linguistiques multi-niveaux, certaines connaissances mono-niveau, appartenant plus précisément au niveau morphologique de la langue, se sont révélées nettement supérieures aux autres pour aider le SRI à retrouver les documents pertinents. Nous cherchons donc à voir s'il est possible d'améliorer encore leur impact. L'apport de la prise en compte d'informations morphologiques en RI n'est certes pas nouveau, de nombreux travaux ayant montré que ces connaissances pouvaient dans certaines conditions être bénéfiques en RI ; mais nous nous penchons sur ce niveau de langue particulier dans le dernier chapitre de ce mémoire en proposant une méthode inédite pour l'acquisition de ces informations morphologiques. Contrairement à la plupart des techniques existantes, notre approche présente la particularité de ne nécessiter aucune connaissance ni donnée externe, d'être entièrement automatique et capable de s'appliquer directement à diverses langues. Le principal objectif de notre dernière contribution est donc de montrer que les informations linguistiques, lorsqu'elles sont extraites à l'aide de techniques simples et qui s'adaptent aux contraintes de la RI, contribuent significativement et de manière plus tranchée à l'amélioration des performances des SRI.

Organisation du manuscrit

Dans le premier chapitre, nous présentons les mécanismes traditionnels de la RI. Nous y dressons un tour d'horizon des principales techniques utilisées par les SRI pour représenter les documents et requêtes, mettre en correspondance leurs contenus et retourner à l'utilisateur les documents dont le contenu est le plus proche de celui de sa requête. Nous tentons également de mettre en évidence les limites de ces mécanismes fondés sur une comparaison de mots et montrons l'intérêt de recourir à des informations plus fines et plus significatives, obtenues par le biais de techniques du traitement automatique des langues. Le chapitre deux propose donc une synthèse des contributions possibles des techniques issues du TAL pour une application en RI à travers un état de l'art des diverses tentatives déjà réalisées dans ce vaste domaine. Nous dressons un bilan de l'apport du couplage TAL-RI, à partir duquel nous pouvons postuler de nouvelles hypothèses et proposer des pistes de recherche pour exploiter autrement les techniques du TAL en RI. Nous cherchons plus particulièrement à évaluer l'intérêt en RI de combiner des informations linguistiques multi-niveaux (d'ordre morphologique, syntaxique et sémantique) plus à même d'exploiter la richesse de la langue. Le chapitre trois présente une architecture pour ce faire et, outre l'analyse de l'apport respectif des diverses connaissances prises en compte, décrit pour répondre à cette question notre analyse originale des corrélations étudiant les relations qu'elles entretiennent dans la recherche de documents pertinents. L'intérêt de coupler des informations multi-niveaux ayant été montré, il est nécessaire de combiner de manière optimale les diverses connaissances prises en compte. Le chapitre quatre présente notre approche de fusion des résultats obtenus par chacune des informations linguistique exploitées par le SRI, approche qui s'adapte d'une part à l'efficacité respective des connaissances multi-niveaux prises en

compte et, d'autre part, aux spécificités des requêtes. Enfin, ces diverses analyses ayant mis en valeur l'intérêt en RI d'exploiter des informations morphologiques, le dernier chapitre décrit la méthode efficace que nous proposons pour leur acquisition, et qui présente la particularité d'être bien adaptée aux contraintes de la RI.

Chapitre 1

Recherche d'information

Résumé : Ce chapitre présente un état de l'art du domaine de la recherche d'information. Il introduit tout d'abord le processus général de recherche d'information, *i.e.* le mécanisme emprunté par les systèmes pour retrouver, parmi une masse de documents, ceux qui répondent précisément au besoin d'information d'un utilisateur. Il décrit ensuite les diverses méthodes de représentation textuelle des documents et requêtes communément utilisées, puis détaille les stratégies adoptées pour permettre la mise en correspondance de ces représentations. Il propose enfin un bref aperçu des techniques d'évaluation traditionnellement utilisées pour juger de la pertinence des systèmes de recherche d'information.

Mots-clés : recherche d'information, indexation, modèles de recherche d'information, pertinence, évaluation des systèmes de recherche d'information.

1.1 Introduction

L'objectif de la recherche d'information (RI) est de concevoir des systèmes (nommés désormais SRI pour systèmes de recherche d'information) capables de retrouver parmi un ensemble de documents ceux qui répondent précisément au besoin d'un utilisateur. Ce besoin est généralement formulé par le biais d'une requête en langage naturel. La principale difficulté des SRI est d'établir un lien qui soit pertinent entre les documents et la requête. Ce chapitre présente les principales techniques et méthodes communément utilisées en RI pour répondre en partie à ce problème de pertinence. En vue de nos travaux de recherche, cet état de l'art se limite exclusivement à l'étude de la RI textuelle et automatique (nommée également recherche documentaire).

En recherche documentaire, les documents et la requête, pour pouvoir être mis en correspondance, doivent tout d'abord être représentés sous une forme qui soit exploitable par un SRI. Pour y parvenir, les systèmes passent généralement par une phase dite d'*indexation* qui a pour objectif d'identifier les idées majeures, les concepts importants des textes ou des questions[1], par une analyse de leurs contenus. L'extraction de concepts

[1] Les termes *question* et *requête* sont employés sans aucune distinction dans la suite de ce mémoire.

étant une opération particulièrement difficile à mettre en œuvre[2], l'approche communément adoptée en RI textuelle est plutôt de chercher des représentants de ces concepts, plus faciles à identifier. Ces représentants correspondent généralement, dans le cadre de l'indexation automatique dans lequel nous nous plaçons, à un ensemble de mots extraits des documents et requêtes, nommés *termes d'indexation*. L'indexation consiste donc à associer à chaque document (ou à chaque requête) un descripteur (également nommé *index*) formé de l'ensemble des termes d'indexation extraits de son contenu. Cette phase d'indexation est cruciale puisque les documents et requêtes sont ensuite uniquement désignés par ces descripteurs. La pertinence des résultats retournés à l'utilisateur est alors fortement dépendante de la qualité de ces représentations. Une partie de ce chapitre y est donc consacrée et présente les approches communément utilisées en RI (section 1.2).

Pour établir une correspondance entre documents et requêtes, représentés par des descripteurs, les SRI s'appuient sur des modèles de RI. Le rôle de ces modèles est triple. Ils permettent :

- de donner une interprétation aux descripteurs en offrant une représentation interne des textes et des questions basée sur les termes d'indexation ;
- de définir les stratégies à adopter pour comparer les représentations des documents et des requêtes. Leur comparaison donne lieu à un score qui traduit leur degré de ressemblance ;
- de proposer éventuellement des méthodes de classement des résultats retournés à l'utilisateur.

Ces modèles sont donc fondamentaux puisqu'ils proposent un cadre théorique pour modéliser la notion de pertinence. Plus les mécanismes mis en œuvre pour apparier les documents et requêtes sont efficaces, plus les documents retournés à l'utilisateur correspondent à l'information recherchée. Nous proposons dans ce chapitre (section 1.3) un tour d'horizon des principaux modèles utilisés en RI, en cherchant à mettre en valeur leurs avantages et leur faiblesses.

Une fois les représentations des documents et des requêtes mises en correspondance, le système retourne à l'utilisateur la liste des documents considérés comme les plus pertinents par rapport à sa requête. Bien que la notion de pertinence d'un document soit, comme nous le verrons, très subjective, des méthodes et des mesures d'évaluation sont nécessaires pour estimer la validité des résultats retournés par le système. Une partie de ce chapitre y est consacrée (section 1.4).

1.2 Indexation et mécanismes fondamentaux de recherche d'information

Après avoir présenté le schéma général du processus de RI et défini ses trois éléments-clés, nous décrivons les méthodes d'indexation des documents et requêtes. Nous détaillons plus précisément les techniques utilisées pour :

- analyser le contenu textuel des documents et requêtes,

[2]Elle nécessite en effet de procéder à une analyse sémantique fine des textes qui fait appel comme nous le verrons dans le deuxième chapitre à des méthodes souvent complexes.

- choisir les termes d'indexation les plus représentatifs des contenus sémantiques des textes et des questions,
- représenter les différents degrés d'importance de ces termes au sein des textes.

Nous décrivons ensuite les mécanismes généraux mis en œuvre afin d'apparier les descripteurs associés aux documents et requêtes et permettre au SRI de retourner à l'utilisateur une liste de documents potentiellement classés par ordre de pertinence. Nous revenons enfin sur une des techniques communément employées pour permettre une description plus précise du besoin d'information de l'utilisateur : les méthodes d'expansion de requête.

1.2.1 Processus général de recherche d'information

Le processus de recherche d'information a pour objectif d'établir une correspondance pertinente entre l'information recherchée par l'utilisateur, représentée généralement par le biais d'une requête, et l'ensemble des documents disponibles. Il s'articule donc autour de trois éléments-clés : le *document*, la *requête* et la notion de *pertinence*. Avant de présenter les grandes étapes du processus de RI, nous proposons de revenir sur la définition de ses trois acteurs principaux.

1.2.1.1 Principaux acteurs du processus

Le document

La notion de document est particulièrement complexe à définir (se référer notamment à (Saracevic, 1996)). Dans son acception courante, l'une des définitions possibles de ce terme est de considérer un document comme le support physique d'une information. Dans le cas des données susceptibles d'être manipulées par un SRI, ce support physique (et plus particulièrement numérique) peut correspondre à un texte (dans son intégralité ou un extrait), une page *Web*, une image, une séquence vidéo ou sonore... Un document est, dans ce cadre, défini comme toute unité susceptible de constituer une réponse à une requête d'un utilisateur. Pour notre part, nous nous intéressons uniquement au document textuel. Un document-texte peut être représenté selon trois vues (Sauvagnat, 2005; Fuhr, 2005) : la vue *présentation* qui décrit la représentation sur un *medium* à deux dimensions (alignement de paragraphes, indentation, en-têtes et pieds de pages...), la vue *logique* qui contient des informations sur la structure et la partition d'un document (*e.g.* une structuration en chapitres, sections...) et la vue *contenu*, appelée également *vue sémantique*, qui se concentre sur le contenu textuel du document, c'est-à-dire l'information qui y est véhiculée. Les représentations logiques sont parfois intégrées en RI dans la notion de document. Une partie de la communauté RI voit en effet dans la structure un moyen d'améliorer la représentation des documents et de localiser plus précisément l'information recherchée. Ces travaux appartenant à la recherche d'information dite structurée (RIS) sont actuellement en plein essor, comme en témoigne la création récente d'une conférence sur l'évaluation des

SRI structurés (*cf.* la campagne d'évaluation INEX[3]). Pour notre part, nous considérons le document uniquement du point de vue de son contenu sémantique (la vue *contenu*). L'unité documentaire est donc représentée par un support physique (le texte) associé à une information véhiculée par son contenu sémantique. L'ensemble des documents mis à disposition du SRI pour lui permettre de retrouver l'information recherchée par l'utilisateur constitue la *base documentaire* (également nommé *fonds documentaire* ou encore *collection de documents*).

La requête

La requête se définit en RI au sein d'un processus cognitif plus large représenté par le *besoin d'information* d'un utilisateur. Ce besoin correspond à l'expression mentale de l'information qu'il recherche. Le passage d'un besoin d'information à son expression en des termes compréhensibles par le SRI est une tâche difficile. Dans l'idéal, l'utilisateur devrait avoir des connaissances sur le système lui-même (*e.g.* connaître les mécanismes de recherche utilisés par le SRI), sur la collection de documents disponibles, sur les thèmes associés à ces documents... Parallèlement, le SRI devrait pouvoir s'adapter au contexte précis du besoin de l'utilisateur, *e.g.* connaître son degré d'expertise dans le domaine de l'information recherchée, ses centres d'intérêts, ses préférences de recherches (type de documents, langue...). Bien que les travaux de recherche en RI portant sur une description plus précise et plus pertinente du besoin d'information soient de plus en plus nombreux (voir notamment les diverses études liées à la personnalisation de l'information ou à la RI basée sur le profil de l'utilisateur), le processus utilisé en RI « traditionnelle » pour passer de ce besoin à une forme directement exploitable par le SRI reste basique. Il s'appuie exclusivement sur l'utilisateur qui exprime son besoin d'information en formulant une requête sous forme de mots-clés ou de phrases en langage naturel. La requête peut donc être considérée comme une description partielle du besoin d'information (les mots-clés étant souvent imprécis et ambigus) à un instant donné. Le caractère évolutif de ce besoin (*i.e.* le fait qu'il peut évoluer au fur et à mesure que l'utilisateur acquiert des éléments d'information supplémentaires) n'est par conséquent pas pris en compte. Les requêtes représentées par un ensemble de mots-clés sont généralement plutôt courtes (pour le cas du *Web* par exemple, la moyenne de mots les composant était de 2,44 pour l'année 2005). Selon le modèle de RI utilisé pour la représentation du contenu des requêtes, ces mots-clés sont parfois reliés à l'aide d'opérateurs booléens (ET, OU, NON...). Les requêtes formulées en langage naturel offrent quant à elles à l'utilisateur la possibilité d'exprimer plus librement ce qu'il recherche. Elles nécessitent généralement, pour pouvoir être représentées et interprétées par le SRI, de faire appel à des traitements linguistiques.

[3] http ://inex.is.informatik.uni-duisburg.de/2006/.

La pertinence

Comme nous l'avons vu, le cœur du problème de la RI réside dans la définition d'une fonction de correspondance entre la requête et l'ensemble des documents disponibles. La pertinence en RI peut être vue sous différents angles (Saracevic, 1996) :

- du point de vue de l'utilisateur ; elle représente alors la façon dont ce dernier évalue les documents retrouvés par le SRI en fonction de son besoin d'information (on parle de ses *jugements de pertinence*). Il s'agit de la pertinence « utilisateur » ;
- du point de vue du système, *i.e.* la pertinence que le système a lui-même calculée à partir des méthodes utilisées pour comparer les documents et la requête. C'est la pertinence « système ».

Rendre compte de manière automatique de la pertinence « utilisateur » semble difficile, notamment compte tenu de son caractère fortement subjectif (un même document peut être jugé pertinent ou non selon les utilisateurs), de son aspect nuancé (deux documents pertinents n'ont pas nécessairement la même importance, et les raisons de cette importance peuvent varier), et évolutif (un document non pertinent à un moment donné peut le devenir au fur et à mesure que l'utilisateur progresse dans sa connaissance du sujet). Cette difficulté est renforcée par le fait que les utilisateurs ont souvent également du mal à définir et à exprimer leur besoin d'information. C'est pourquoi, en RI, la notion de pertinence est représentée essentiellement au travers de la pertinence « système ». Cette dernière s'exprime sous la forme d'un score obtenu automatiquement par les SRI en comparant les représentations des documents et celles des requêtes suivant les méthodes définies par le modèle de RI utilisé. Bien que ce score ne soit qu'une représentation imprécise de la pertinence « utilisateur » (un document considéré comme pertinent par le système ne l'est pas nécessairement par l'utilisateur), il présente l'avantage de rendre « mesurable » la notion de pertinence et de permettre par conséquent l'évaluation des performances des SRI. Tout l'enjeu de la RI réside dans la mise en œuvre de mécanismes visant à rapprocher la pertinence système de la pertinence utilisateur.

Nous venons de définir les trois acteurs-clés de tout système de recherche d'information. Nous proposons de revenir dans la section suivante sur le processus de recherche emprunté par ces systèmes pour faire le lien entre ces trois acteurs.

1.2.1.2 Description du processus de RI

Pour mettre en correspondance les requêtes et les documents, un système de recherche d'information s'appuie sur un certain nombre de processus, articulés autour de deux étapes essentielles : les phases d'*indexation* et de *recherche*. Le processus complet de recherche documentaire est représenté en figure 1.1.

L'étape d'*indexation* consiste à analyser les documents et les requêtes afin de créer une représentation de leur contenu textuel qui soit exploitable par le SRI. Chaque document (et requête) est alors associé à un descripteur représenté par l'ensemble des termes d'indexation extraits.

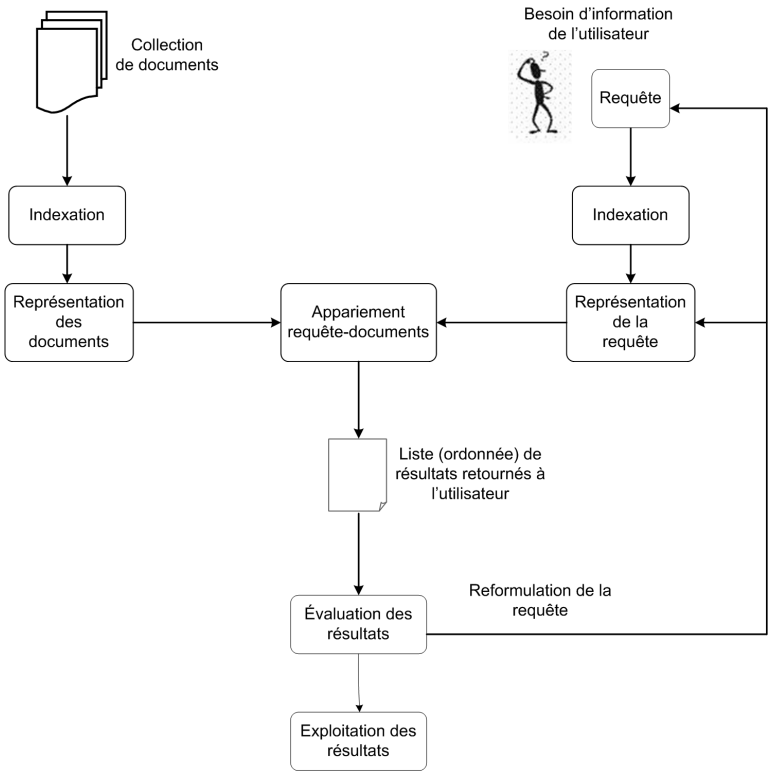

FIG. 1.1 – Processus en U de recherche d'information

La phase de *recherche* vise à apparier les documents et la requête de l'utilisateur en comparant leurs descripteurs respectifs. Pour cela, elle s'appuie sur un formalisme précis défini par un modèle de RI. Les documents présentés en résultat à l'utilisateur, et considérés comme les plus pertinents, sont ceux dont les termes d'indexation sont les plus proches de ceux de la requête.

Tout processus de recherche d'information passe donc nécessairement par ces deux phases. Comme évoqué précédemment, une étape de *reformulation* peut venir compléter ce processus dans le but d'améliorer les mécanismes de recherche en tentant de rapprocher la requête de l'utilisateur de son besoin d'information initial.

La performance d'un SRI est étroitement liée aux méthodes mises en œuvre à la fois pour la représentation des documents et requêtes, et pour leur appariement. Les deux sections suivantes décrivent successivement plus en détail les techniques communément utilisées lors de ces deux phases.

1.2.2 Indexation des documents et requêtes

L'indexation automatique consiste, nous l'avons vu, à extraire des documents et requêtes l'ensemble des mots susceptibles de représenter au mieux leurs contenus sé-

mantiques. La sélection de ces termes d'indexation nécessite toutefois de répondre aux trois questions suivantes :

- qu'entend-on exactement par « mot » ? Comment les identifier au sein des textes ?
- comment choisir, parmi tous les mots d'un texte ou d'une question, ceux qui sont les plus significatifs ?
- comment représenter leurs différents degrés de représentativité du contenu textuel ?

Les sections suivantes s'attachent à décrire les méthodes généralement utilisées en RI pour répondre à ces questions.

1.2.2.1 Reconnaissance des mots

La première étape de représentation des documents et requêtes consiste à segmenter le texte en une suite d'unités lexicales élémentaires, prenant classiquement en RI la forme de mots. Le traitement associé est la *tokenisation*. Cette phase est beaucoup moins évidente qu'il n'y paraît, la notion de *mot* étant particulièrement complexe à définir. Une des définitions possibles consiste à considérer un mot comme une suite de caractères graphiques (pour la langue écrite) formant une unité sémantique et pouvant être distinguée par des séparateurs. La plupart des SRI s'appuie sur cette définition pour leur segmentation. La principale difficulté de ce traitement réside dans la sélection des délimiteurs utilisés pour établir les frontières entre mots. Ceux-ci, généralement représentés par le blanc typographique et les signes de ponctuation, sont dans certains cas problématiques, comme l'illustrent les différentes unités suivantes : *aujourd'hui*, *3h*, *après-midi*, *Mr J.-P. Martin*, *parce que*, *l'O.N.U*... Ces difficultés sont plus ou moins importantes selon les langues prises en compte. Pour les langues germaniques comme l'allemand par exemple, il est nécessaire de prendre en considération leur nature fortement agglutinante (*i.e.* la présence de plusieurs mots au sein d'une même chaîne de caractères). Pour de nombreuses langues asiatiques, la tâche est encore plus ardue, celles-ci ne disposant d'aucun délimiteur visible entre mots. Certains SRI s'affranchissent de ce traitement de *tokenisation* et s'appuient, pour extraire les unités textuelles, sur des modèles de n-grammes. Le texte est considéré comme une suite de N caractères consécutifs. Pour une position i (où $1 \leq i \leq N$) dans le texte, on sélectionne une suite de n caractères consécutifs. Le texte est alors représenté par l'ensemble de ces « extraits ».

D'une manière générale, la notion de mot (considéré comme une suite de caractères comprise entre deux séparateurs) telle qu'elle est prise en compte par la majorité des SRI est très réductrice de la richesse de la langue. En effet, elle ne permet pas de rendre compte par exemple du phénomène de variation linguistique (*i.e.* considérer par exemple que les quatre mots *beau*, *belle*, *beaux* et *belles* correspondent en fait à une seule et même unité lexicale), ni du fait qu'il existe des relations entre les termes (*i.e.* considérer par exemple que les éléments de l'expression *effet de serre* forment une unité de sens à part entière, dont la signification dépasse celle de ses éléments pris isolément). Une solution souvent adoptée pour représenter le contenu textuel par des unités plus fines que de simples chaînes de caractères consiste à appliquer une analyse linguistique des textes et des questions, qui permet de considérer les mots comme des

entités linguistiques à part entière. Le deuxième chapitre de ce mémoire revient plus en détail sur l'apport de ces traitements linguistiques en RI.

Les documents et requêtes ayant été segmentés en une suite de mots, il convient alors de sélectionner ceux qui sont les plus susceptibles d'être de bons candidats pour représenter leurs contenus sémantiques respectifs. Cette étape fait l'objet de la section suivante.

1.2.2.2 Sélection des termes d'indexation

La phase de sélection des termes d'indexation est fondamentale dans le processus de RI : ce sont en effet ces termes qui vont représenter les documents et requêtes au sein du SRI. Il convient donc de choisir ceux qui reflètent le mieux leur contenu sémantique. Cette sélection est généralement liée à une phase de pondération décrite en section suivante. Dans l'idéal, les termes retenus doivent, d'une part, être le plus univoque et discriminant possible et, d'autre part, être en nombre limité afin de ne pas complexifier les calculs effectués lors de la comparaison des représentations. Plusieurs traitements complémentaires peuvent être utilisés par les SRI pour pouvoir répondre à ces deux exigences.

Élimination des mots-vides

Le premier traitement consiste à supprimer les mots dont on sait par avance qu'ils sont peu informatifs. Ce sont généralement des mots dits « grammaticaux » (comme les prépositions *à*, *de*, les articles *le*, *la*, *un*, *des*, les pronoms *ce*, *lui* ou encore les auxiliaires *être*, *avoir*...), ou des mots très fréquents au sein d'une collection de textes donnée (par exemple, le mot *informatique* dans un corpus spécialisé dans ce domaine). L'élimination de ces mots, nommés le plus souvent *mots-vides*, se fait par le biais d'anti-dictionnaires (ou *stop-lists*) qui recensent l'ensemble des mots d'une langue considérés comme non pertinents pour l'indexation. Ces listes sont utilisables d'une collection à une autre et peuvent être complétées par les mots courants spécifiques au domaine étudié.

Analyse basée sur les fréquences d'occurrences des mots

Le second traitement consiste à choisir les termes d'indexation en fonction de leur fréquence d'apparition dans les textes. Il s'appuie sur des méthodes numériques qui trouvent principalement leurs origines dans la loi de Zipf (1949) et la conjecture de Luhn (1978).

Les travaux de Zipf figurent parmi les premiers à avoir décrit la répartition statistique des fréquences d'apparition des mots au sein des textes et constaté des régularités. Ils montrent que si les termes sont rangés par ordre décroissant de leur fréquence d'apparition au sein d'un texte (ou d'une collection), il existe alors une relation entre le rang de ces termes et leur fréquence. Cette relation (représentée par l'hyperbole dans

la figure 1.2) peut s'exprimer par la formule suivante :

$$rang * (fréquence\ du\ terme/nombre\ de\ termes) = constante$$

qui signifie que si le rang d'un mot est multiplié par le nombre de fois où il apparaît dans les textes, on aura tendance à trouver un nombre constant. Par exemple, si le mot le plus fréquent d'un texte (rang = 1) apparaît 1000 fois, le deuxième mot aura tendance à se trouver 500 fois dans le texte et ainsi de suite... À la fin de cette liste, on trouvera 1000 mots n'ayant été utilisés qu'une seule fois dans le texte. La loi de Zipf est l'une des premières à avoir montré que les mots dans les documents ne s'organisent pas de manière aléatoire.

Il a également été démontré, notamment dans (Rijsbergen, 1979), que dans un texte, la valeur informative d'un mot peut s'exprimer sous la forme d'une gaussienne en fonction du rang des termes d'un document. Cette courbe, illustrée en figure 1.2, montre que les termes les plus informatifs ne sont ni ceux qui ont une fréquence élevée (*e.g.* les mots-vides) ni ceux qui apparaissent très peu (les mots mal orthographiés ou les néologismes par exemple). La conjecture de Luhn (1978) s'appuie sur cette observation pour spécifier des seuils (correspondant aux seuils *min* et *max* sur la figure 1.2) qui déterminent le pouvoir d'expression des termes. Les mots situés au-delà du seuil maximum sont considérés comme trop communs et ceux en deçà du seuil minimal comme trop rares. Un terme qui se situe entre ces deux extrêmes a par conséquent de forte chance d'être représentatif du contenu informationnel. Ces seuils sont dépendants de la collection utilisée. Il est donc généralement nécessaire de procéder par essais successifs pour trouver leurs valeurs optimales. Pour limiter l'intervention humaine, Salton (1975), après plusieurs expérimentations, propose de considérer comme termes à bon pouvoir discriminant les mots ayant une fréquence en documents comprise dans l'intervalle $\left[\frac{|C|}{100}, \frac{|C|}{10}\right]$, où $|C|$ correspond au nombre de mots dans la collection. Cette conjecture, qui présente également l'avantage de réduire le nombre de mots à utiliser pour la description des textes, permet de sélectionner les termes qui sont considérés comme représentatifs du contenu informationnel des documents et requêtes.

L'étape suivante consiste à attribuer à chacun des termes retenus un poids en fonction de son degré de représentativité. Nous explicitons à présent les mesures de pondération traditionnellement utilisées en RI.

1.2.2.3 Pondération des termes

Une fois les termes d'indexation choisis, il est possible de préciser grâce à un poids que tel terme est plus important que tel autre pour décrire le document ou la requête. Bien que cette pondération soit fortement dépendante du modèle de RI utilisé pour l'appariement de la requête et du document (se référer à la section 1.3), la plupart des méthodes utilisées s'appuient généralement sur la combinaison de trois facteurs : un facteur de pondération locale qui quantifie l'importance du terme dans le document, un facteur de pondération globale qui mesure la représentativité du terme dans l'ensemble de la collection de documents, et un facteur de normalisation qui prend en considération

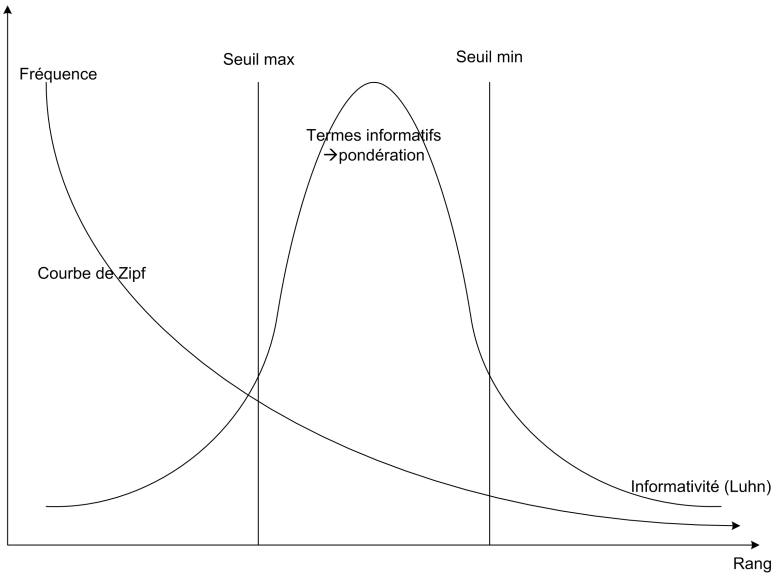

FIG. 1.2 – Relation entre la fréquence et le rang d'un terme (loi de Zipf et conjecture de Luhn) (figure inspirée de (Nie, 2006))

la taille du document (Salton et Buckley, 1988; Singhal, 1997; Robertson et Walker, 1994).

Pondération locale

La pondération locale d'un terme mesure son importance au sein d'un document. Ce type de pondération prend en compte les informations locales du terme qui ne dépendent que du document. Elle correspond généralement à une fonction de la fréquence d'apparition du terme dans le document, notée *tf* (*terme frequency*). Un terme qui apparaît fréquemment dans un document sera considéré comme pertinent pour décrire son contenu.

Les mesures les plus utilisées pour calculer la pondération locale sont les suivantes :

- le *facteur binaire* : il prend pour valeur 1 si le terme est présent dans le document (quelle que soit sa fréquence) et 0 s'il est absent. Ce facteur est généralement utilisé pour les représentations de type ensembliste (*cf.* section 1.3.1) ;
- le *facteur de fréquence tf* : il indique le nombre d'occurrences d'un terme donné dans le document ;
- le *facteur de fréquence normalisée* : il prend en compte la fréquence d'un terme dans le document et mesure son importance relativement aux autres termes du document. Il s'écrit :

$$w(t_i, d_j) = \frac{tf(t_i)}{max_{t \in d_j} tf(t)}$$

où $w(t_i, d_j)$ représente le poids du terme t_i dans le document d_j, $tf(t_i)$ est la fréquence du terme t_i dans le document d_j et $max_{t \in d_j} tf(t)$ est la fréquence maximale d'un terme t dans le document d_j ;

- le *facteur logarithmique* : ce facteur est une fonction logarithmique de la fréquence du terme dans le document, valant :

$$w(t_i, d_j) = \begin{cases} 1 + log(tf(t_i)) \text{ si } tf(t_i) > 0 \\ 0 \text{ sinon} \end{cases}$$

Cette mesure, proposée par Buckley *et al.* (1992), permet de ne pas accorder plus d'importance à un document qui possède un grand nombre de fois un des termes de la requête par rapport à un document qui contient peu de fois plusieurs termes de la requête ;

- le *facteur augmenté* (Salton et Buckley, 1988) : il permet de réduire les différences entre les valeurs associées aux termes du document. Il attribue un poids minimal (généralement de 0,5) aux termes présents dans le document et ne dépassant pas une valeur maximale pour les termes présents plusieurs fois. Il est représenté par la formule suivante :

$$w(t_i, d_j) = \begin{cases} 0,5 + 0,5 \frac{tf(t_i)}{max_{t \in d_j} tf(t)} \text{ si } tf(t_i) > 0 \\ 0 \text{ sinon} \end{cases}$$

Pondération globale

La pondération globale mesure l'importance d'un terme au sein de l'ensemble des documents de la collection. Elle vise à représenter son caractère discriminant, *i.e.* sa capacité à différencier les documents de la collection. Un terme apparaissant dans peu de documents est considéré comme plus discriminant et doit être privilégié par rapport à un terme présent dans de nombreux documents. Plusieurs études ont souligné l'intérêt en RI de prendre en compte l'importance d'un terme dans la collection (Salton *et al.*, 1975). Le calcul de la pondération globale s'appuie notamment sur le nombre de documents de la collection dans lesquels un terme apparaît. L'une des mesures les plus utilisées est la fréquence documentaire inverse, notée *idf* (*inverse document frequency*) (Salton *et al.*, 1975), représentée par la formule suivante :

$$idf(t_i) = log \frac{N}{n_i}$$

où n_i est le nombre de documents contenant le terme t_i et N le nombre total de documents dans la collection. Une variante de cette mesure (nommée fréquence documentaire inverse probabiliste) peut également être utilisée, et est notée :

$$idf\_prob(t_i) = log \frac{N - n_i}{n_i}$$

.

Normalisation

Pour être considéré comme important, un terme doit à la fois être très présent dans le document (pondération locale élevée) et discriminant par rapport aux autres termes de la collection (pondération globale importante). Il est toutefois nécessaire de prendre également en compte un autre critère, celui de la longueur des documents qui joue un rôle important lors de la comparaison des documents et requêtes. En effet, un document très long aura tendance à utiliser les mêmes termes de façon répétée, et sera donc favorisé si la requête porte sur l'un de ces termes. Inversement, un document court, pouvant être pertinent pour une requête donnée, contiendra peu de mots et donc peu de termes avec un poids important. Afin de contourner ce problème, plusieurs facteurs de normalisation ont été proposés (Robertson et Walker, 1994; Singhal, 1997) pour intégrer la taille des documents aux formules de pondération. Parmi ceux-ci, on trouve les facteurs de normalisation L_1 et L_2 (normalisation par le cosinus) de valeurs respectives :

$$\frac{1}{\sum_{i=1}^{n}(l_i.g_i)} \text{ et } \frac{1}{\sqrt{\sum_{i=1}^{n}(l_i.g_i)^2}}$$

où l_i et g_i correspondent respectivement à la pondération locale et globale du $i^{ième}$ terme et n représente le nombre de termes d'indexation.

Combinaison des pondérations

Finalement, le poids d'un terme dans un document, reflétant son pouvoir de représentativité du contenu textuel, est obtenu par la combinaison des pondérations locale (notée l_i) et globale (g_i) et de la normalisation (n_i), *i.e.* : $w(t_i, d_j) = l_i \times g_i \times n_i$.

Nous avons présenté quelques-unes des principales méthodes de pondération utilisées par les SRI lors de l'indexation automatique des documents et requêtes. Cette description n'est cependant pas exhaustive, le choix des mesures à appliquer étant fortement lié au modèle de RI utilisé. En résultat de l'indexation, chaque document (et chaque requête) est donc associé à un ensemble de termes pondérés, son descripteur, représentatif de son contenu. Nous proposons à présent de décrire comment, à partir de ces descripteurs, le SRI établit une correspondance entre la requête et les documents afin de retourner à l'utilisateur une liste de documents pertinents étant donné son besoin d'information.

1.2.3 Processus de recherche des documents pertinents

Le processus de recherche est fondamental en RI puisque c'est lui qui permet aux SRI d'établir un lien entre les documents de la collection et la requête. Cette étape étant étroitement liée aux spécificités des modèles de RI utilisés par ces systèmes (présentés en section 1.3), nous n'en détaillons ici que les principes fondamentaux.

L'objectif du processus de recherche est de mesurer la pertinence d'un document par rapport à une requête. Pour y parvenir, il s'agit tout d'abord de donner une interprétation aux descripteurs obtenus lors de la phase d'indexation. Pour cela, on crée dans un premier temps, en s'appuyant sur le formalisme défini par le modèle de RI, une représentation interne des textes et des questions à partir de leurs termes d'indexation. Ces représentations, utilisant un formalisme identique, peuvent alors être comparées les unes aux autres. Le résultat de cette comparaison se traduit par un score qui détermine le degré de pertinence du document par rapport à la requête. La pertinence est celle du système, et doit être la plus proche possible de la pertinence « utilisateur ». Plus précisément, le score est calculé à partir d'une fonction de similarité notée *RSV(d,q)* (*RSV* pour *retrieval status value*), où d représente un document et q la requête de l'utilisateur. Cette fonction définie par le modèle de RI tient compte de la pondération attribuée aux termes lors de l'étape d'indexation. Chaque document étant associé à un score de pertinence, il est alors possible d'établir un classement ordonné des documents (les premiers documents étant ceux qui possèdent la représentation la plus proche de celle de la requête), éventuellement grossier selon les modèles. Ce classement permet de constituer la liste des résultats retournés à l'utilisateur. Quel que soit le système utilisé, et aussi performant qu'il puisse être, cette liste ne représente néanmoins qu'une partie de l'ensemble des documents pertinents effectivement disponibles au sein de la collection. Lors des dernières campagnes d'évaluation TREC, les SRI les plus performants atteignaient ainsi rarement 40% du total des réponses attendues. Compte tenu de ce constat, il est nécessaire de mettre en œuvre d'autres techniques, en complément du processus de recherche, destinées à permettre aux SRI de retrouver davantage de documents pertinents. Nous terminons cette section consacrée à la description des mécanismes traditionnels de RI en présentant l'une d'elles : la reformulation de la requête de l'utilisateur.

1.2.4 Phase de reformulation

Sans une connaissance approfondie de la collection de documents et des mécanismes de recherche précis utilisés par le SRI, il est difficile pour la plupart des utilisateurs de formuler la requête « idéale » qui va permettre de retrouver l'information exacte recherchée. Cette observation est d'autant plus vraie dans le cas des moteurs de recherche sur le *Web* où les utilisateurs passent beaucoup de temps à reformuler leur requête pour trouver les documents recherchés. La recherche d'informations pertinentes à partir de la seule requête initiale, généralement limitée à peu de mots, est une tâche très difficile à réaliser (Rijsbergen, 1986). C'est pourquoi, une étape de reformulation automatique de la requête est souvent intégrée dans le mécanisme de RI. Elle consiste à modifier la requête initiale, principalement en ajoutant de nouveaux termes susceptibles de représenter plus précisément le besoin d'information, et en ré-estimant le poids des termes initiaux.

Différents mécanismes ont été proposés en RI pour ce faire. Nous abordons ici essentiellement les méthodes d'expansion de requêtes. Nous distinguons plus précisément les techniques qui s'appuient sur des ressources de celles qui utilisent uniquement des

informations issues des documents et requêtes (les techniques dites de « rétroaction » de pertinence (*relevance feedback*)).

Méthodes d'expansion de requêtes basées sur des ressources

La stratégie généralement adoptée pour la reformulation de la requête consiste à enrichir cette dernière à l'aide de connaissances complémentaires issues de ressources. Ces informations visent à préciser la question (en identifiant par exemple le sens de ses termes) ou à l'élargir (à l'aide de mots sémantiquement proches mais différents (des synonymes par exemple)). Elles peuvent être obtenues à partir :

- de ressources externes, telles que des bases de connaissances linguistiques (dictionnaires, thésaurus...) ;
- de ressources internes : les informations sont acquises directement à partir de la collection de documents (*e.g.* cooccurrences).

Ces méthodes font généralement appel à des informations issues d'une analyse linguistique des documents et requêtes. L'objectif de ce chapitre étant de présenter uniquement les méthodes traditionnelles de RI (*i.e.* sans apport de traitements linguistiques), elles seront par conséquent présentées plus en détail au chapitre 2. D'une manière générale, le principal bilan que l'on peut faire est que l'enrichissement des requêtes par le biais de ressources (externes ou internes) est efficace uniquement si les mots ajoutés sont véritablement liés sémantiquement aux constituants de la question.

Techniques de *relevance feedback*

De nombreux travaux de recherche font de la reformulation de requêtes grâce à une technique particulière : le *relevance feedback*[4]. Cette technique consiste, de manière simplifiée, à utiliser la requête initiale pour amorcer la recherche puis, en s'appuyant sur les documents pertinents obtenus, à modifier celle-ci en lui ajoutant des termes extraits de ces documents pertinents. La nouvelle requête ainsi obtenue permet de se rapprocher progressivement — ce processus pouvant être répété plusieurs fois — du besoin d'information de l'utilisateur.

Le processus de *relevance feedback* s'articule donc autour de deux éléments-clés : la sélection des documents pertinents obtenus lors d'une première recherche, et le choix des termes à ajouter à la requête. Certaines techniques de *relevance feedback* font appel à l'utilisateur (Rocchio, 1971) : ce dernier intervient pour sélectionner les documents (obtenus lors de la première recherche) qui lui semblent être les plus pertinents. D'autres sont entièrement automatiques (Salton et Buckley, 1990) ; on parle de *pseudo-relevance feedback*. Ce sont les n premiers documents de la liste de résultats obtenue lors du premier passage qui sont alors considérés comme pertinents. Les méthodes utilisées ensuite pour l'extraction des nouveaux termes sont variées — elles peuvent notamment s'appuyer sur la fréquence de ces termes au sein des documents

[4]Pour un état de l'art détaillé, se référer à (Ruthven et Lalmas, 2003; Salton et Buckley, 1990; Lundquist *et al.*, 1997) *inter alia*.

pertinents — et sont étroitement liées au modèle de RI utilisé (*cf.* notamment (Harman, 1992; Salton et Buckley, 1990) pour le modèle vectoriel et (Rijsbergen, 1977; Harman, 1992; Robertson *et al.*, 1981) pour le modèle probabiliste). De nombreux travaux ont montré l'apport significatif de ces techniques de reformulation en RI (Salton et Buckley, 1990; Ruthven et Lalmas, 2003).

Nous venons de présenter entre autres deux phases fondamentales (indexation et recherche) du processus de RI, empruntées par les SRI pour fournir à l'utilisateur un ensemble de documents susceptibles de répondre précisément à son besoin d'information initial. Bien que les mécanismes mis en œuvre soient communs à tous ces systèmes, les méthodes employées sont très différentes et fortement liées au modèle de RI utilisé. La section suivante présente les principaux modèles à la base des SRI actuels.

1.3 Modèles de RI

Comme nous l'avons vu, tout l'art d'un SRI réside dans sa capacité à établir la pertinence d'un document vis-à-vis d'une requête. Pour parvenir à cet objectif, il s'appuie sur un modèle de RI, dont le principal rôle est de proposer un cadre théorique solide pour représenter cette notion de pertinence. D'une manière formelle, un modèle de RI peut être défini (Baeza-Yates et Ribeiro-Neto, 1999) par le quadruplet suivant {d, q, $\Im$, RSV(d,q)} où :

- d et q correspondent respectivement à la représentation d'un document et d'une requête de la collection (obtenue à la suite de l'indexation),
- $\Im$ est le formalisme d'expression interne des représentations de d et q et leurs relations,
- $RSV(d,q)$ est la fonction de correspondance (*ranking*).

De nombreux modèles ont été proposés en RI. En fonction du cadre théorique sur lequel ils s'appuient, ils sont généralement regroupés autour des trois familles suivantes :

- les modèles *ensemblistes* qui considèrent le processus de recherche comme une succession d'opérations à effectuer sur des ensembles de mots contenus dans les documents,
- les modèles *algébriques* au sein desquels la pertinence d'un document par rapport à une requête est envisagée à partir de mesures de distance dans un espace vectoriel,
- les modèles *probabilistes* qui représentent la RI comme un processus incertain et imprécis où la notion de pertinence peut être vue comme une probabilité de pertinence.

Le schéma en figure 1.3, inspiré de (Baeza-Yates et Ribeiro-Neto, 1999), présente la répartition des principaux modèles existants autour de ces trois familles.

Les sections suivantes s'attachent à proposer une description succincte des modèles les plus fréquemment utilisés en RI. Notre objectif ici est de présenter leurs principes généraux, *i.e.* le formalisme proposé pour créer une représentation interne des documents et requêtes, les stratégies d'appariement mises en œuvre pour comparer ces représenta-

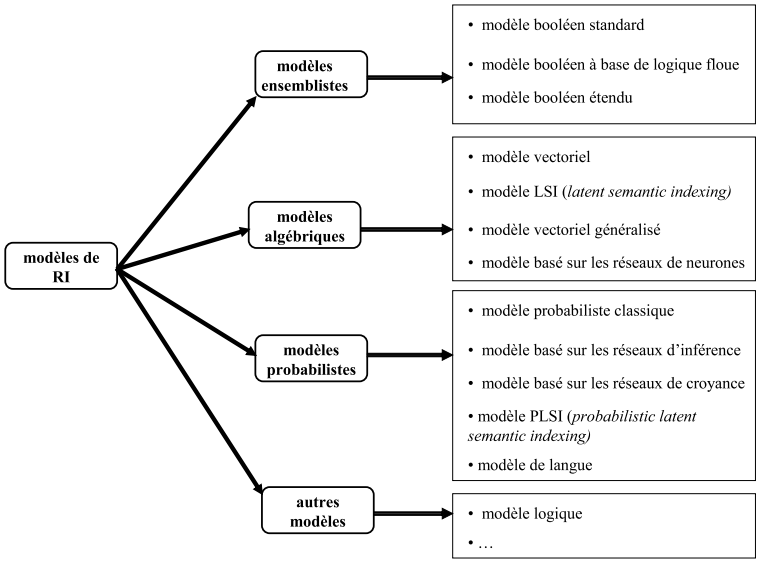

FIG. 1.3 – Taxonomie des principaux modèles de RI

tions, et les fonctions de correspondance utilisées pour associer un score de pertinence au document. Cette présentation ne prétend pas à l'exhaustivité. Un panorama plus complet peut être trouvé dans (Baeza-Yates et Ribeiro-Neto, 1999) ; des références détaillant plus précisément chacun de ces modèles sont également proposées au fil du texte.

1.3.1 Modèles ensemblistes

Nous nous intéressons ici uniquement au principal représentant des modèles inspirés de la logique booléenne et de la théorie des ensembles pour modéliser l'appariement entre une requête et les documents de la collection : le modèle booléen classique.

Le modèle booléen est l'un des plus simples et des plus anciens de la recherche documentaire. Dans ce modèle, un document est représenté comme la conjonction de l'ensemble des termes qui le composent. Une requête est quant à elle considérée comme une expression logique dont les termes sont reliés par les opérateurs de conjonction ($\wedge$), de disjonction ($\vee$) ou de négation ($\neg$). La pertinence entre le document d et la requête q (notée RSV(d,q)) se calcule alors récursivement de la manière suivante :

– si la requête contient un seul terme, on a $q = t$ (avec t un terme) :

$$RSV(d,q) = 1 \; si \; t \in d, \; 0 \; sinon$$

– si la requête contient deux termes reliés par l'opérateur $\wedge$, on a $q = t_1 \wedge t_2$:

$$RSV(d,q) = 1 \; si \; RSV(d,t_1) = 1 \; \textbf{et} \; RSV(d,t_2) = 1, \; 0 \; sinon$$

– si la requête contient deux termes reliés par l'opérateur $\vee$, on a $q = t_1 \vee t_2$:

$$RSV(d,q) = 1 \; si \; RSV(d,t_1) = 1 \; \textbf{ou} \; RSV(d,t_2) = 1, \; 0 \; sinon$$

– si la requête est composée de la négation d'un seul terme, on a $q = \neg t$:

$$RSV(d,q) = 1 \; si \; t \notin d, \; 0 \; sinon$$

Ainsi, une requête : *(orange* $\wedge$ *(ville* $\vee$ *cité))* $\wedge$ *(*$\neg$ *(réseau* $\vee$ *opérateur* $\vee$ *mobile))* retourne à l'utilisateur les documents contenant obligatoirement le terme *orange* et l'un des deux termes *ville* ou *cité* mais qui ne contiennent en aucun cas les termes *réseau*, *opérateur* et *mobile*.

Bien qu'il présente l'avantage d'être simple à appréhender et puisse s'avérer efficace si l'utilisateur maîtrise parfaitement le langage de requêtes, le modèle booléen connaît plusieurs limites. Il propose une modélisation assez « pauvre » de la notion de pertinence. Cette dernière repose en effet sur un critère exclusivement binaire : un document est soit pertinent, soit non pertinent. Il ne prend pas non plus en considération la pondération des termes : un mot a un poids égal à 1 s'il appartient au document, 0 sinon. De ces problèmes découlent deux principales conséquences. Les résultats retournés à l'utilisateur ne peuvent être classés : les textes ramenés ont tous la même importance, et les documents qui ne contiennent pas tous les termes de la requête sont automatiquement considérés comme non pertinents (une requête composée des termes t_1, t_2 et t_3 ne pourra pas par exemple être appariée avec un document composé uniquement des termes t_1 et t_2). Les SRI basés sur ce modèle ont par conséquent de grandes chances de passer à côté de documents susceptibles pourtant d'intéresser l'utilisateur.

Compte tenu de ces limites, il semble indispensable de donner un peu de souplesse à ce modèle. Les adaptations proposées dans le cadre des modèles booléens étendus (*P-norm model* (Salton *et al.*, 1983)) — qui proposent de considérer les opérations booléennes en termes de distances algébriques — ou à base de logique floue (Kraft et Buell, 1983; Ogawa *et al.*, 1991; Dubois *et al.*, 1997) — qui intègrent le poids des termes dans les documents — vont dans cette direction.

1.3.2 Modèles algébriques

Les modèles algébriques rassemblent les modèles de RI qui proposent une représentation vectorielle des documents et requêtes. À partir de ces représentations, la mise en correspondance d'un document et d'une requête revient à appliquer un calcul algébrique de similarité entre vecteurs. Parmi les nombreux modèles qui s'appuient sur ce cadre théorique émergent le modèle vectoriel, certainement le plus répandu, mais également des modèles plus complexes comme les réseaux de neurones (Baeza-Yates et Ribeiro-Neto, 1999), les modèles vectoriels généralisés (Wong *et al.*, 1985) et les modèles LSI (*Latent Semantic Indexing*) (Deerwester *et al.*, 1990). Les travaux présentés dans le cadre de cette thèse s'appuyant essentiellement sur le modèle vectoriel, nous en proposons une description assez détaillée. Les autres modèles ne sont ici que brièvement évoqués.

1.3.2.1 Modèle vectoriel

Le modèle vectoriel (nommé également *VSM* pour *Vector Space Model*) a été proposé par Salton en 1971 (Salton, 1971). Sa forme d'implémentation la plus connue est le système de recherche documentaire SMART (*Salton's Magical Automatic Retriever of Text*), encore utilisé aujourd'hui. Ce modèle représente les requêtes et les documents sous forme de vecteurs qui sont placés dans un même espace vectoriel. Les documents considérés comme les plus pertinents sont ceux dont le vecteur est le plus proche de celui de la requête, suivant une mesure de similarité définie au préalable.

D'une manière plus précise, un vecteur document (ou un vecteur requête) est composé de caractéristiques. Une caractéristique correspond généralement à un terme du document (terme d'indexation), et est associée à un poids qui représente l'importance de ce terme dans le document. Un document (ou une requête) est donc représenté par un vecteur de termes pondérés dans un espace à n dimensions, où n représente le nombre des termes d'indexation de la collection. Un vecteur document $\vec{d_j}$ et un vecteur requête $\vec{q}$ sont donc définis de la manière suivante :

$$\vec{d_j} = (w_{j,1}, \ldots, w_{j,i}, \ldots, w_{j,n})$$

$$\vec{q} = (w_{q,1}, w_{q,i}, \ldots, w_{q,n})$$

où $w_{j,i}$ (resp. $w_{q,i}$) représente le poids du $i^{ième}$ terme dans le document j (resp. dans la requête q).

Les pondérations accordées aux termes des documents (et de la requête) (Salton et Buckley, 1988) prennent généralement en compte les facteurs combinés de pondération locale, globale et de normalisation, tels qu'ils ont été présentés en section 1.2.2.3. Le mécanisme d'appariement consiste alors à évaluer la similarité entre les vecteurs des documents et le vecteur requête. Parmi les différentes mesures existantes (Baeza-Yates et Ribeiro-Neto, 1999), une des plus utilisées est la mesure du cosinus de l'angle entre les deux vecteurs de termes pondérés, notée :

$$sim(\vec{d_j}, \vec{q}) = \frac{\sum_{i=1}^{n} w_{j,i} \times w_{q,i}}{\sqrt{\sum_{i=1}^{n} w_{j,i}^2} \times \sqrt{\sum_{i=1}^{n} w_{q,i}^2}}$$

La valeur associée à la mesure de l'angle formé entre les deux vecteurs correspond au score qui sera attribué à chaque document, et traduit son degré de pertinence par rapport à la requête de l'utilisateur. D'autres mesures sont parfois employées (pour une description de ces mesures, se référer à (Besançon, 2002)). Au final, les documents retournés à l'utilisateur sont classés dans l'ordre décroissant de leur degré de similarité avec la requête.

Les principaux avantages du modèle vectoriel résident dans sa simplicité, sa robustesse et sa rapidité. Les nombreux critères qu'il offre pour la pondération des termes en font son principal atout. Les mesures de similarité utilisées permettent d'ajouter à la notion de pertinence un degré d'« approximation ». Un document peut ainsi être considéré comme pertinent même s'il ne contient pas tous les termes de la requête. Le classement ordonné des résultats constitue également l'un des points forts de ce modèle.

L'une de ses limites est liée à sa façon de représenter le contenu des documents et requêtes. Cette représentation dite en « sac de mots » (un document est transformé en un vecteur de termes éventuellement pondérés) présente en effet l'inconvénient majeur de ne pas prendre en compte l'ordre des mots. Ainsi, pour les deux requêtes : *la voile du bateau* et *le bateau à voile*, ce modèle offre la même représentation interne (hors pondération) : $\vec{q} = (bateau, voile)$. Un autre problème de ce modèle concerne l'hypothèse d'orthogonalité qu'il fait des dimensions de l'espace vectoriel et qui présuppose, puisque les dimensions sont représentées par les termes, que ces derniers sont indépendants les uns des autres (Raghavan et Wong, 1986). Or, cette hypothèse n'est pas valide dans la pratique puisque la plupart des mots d'une langue entretiennent les uns avec les autres des relations de natures diverses. Avec ce type de représentation, un document qui contiendrait par exemple le terme *automobile* et une requête représentée par le terme *voiture* ne peuvent être appariés, chaque terme étant représenté par une dimension différente. Ces deux limites attestent de l'idée que les modèles vectoriels (et la plupart des modèles de RI présentés ici) sont peu adaptés à accueillir des informations plus riches que de simples mots. Or, cette observation est en contradiction avec l'objectif même que nous nous sommes fixé dans le cadre de thèse : exploiter la richesse des informations linguistiques pour améliorer le processus de RI. Nous reviendrons donc régulièrement sur cette problématique au fil de ce mémoire.

Des variantes du modèle vectoriel ont toutefois été proposées pour tenter de contourner ces limites. Nous pouvons notamment citer le modèle vectoriel généralisé (nommé *GVSM* pour *Generalized Vector Space Model* (Wong *et al.*, 1985)) ou le modèle LSI — décrit brièvement dans la section suivante — qui visent à relâcher l'hypothèse d'indépendance entre les termes en prenant en compte leurs cooccurrences.

1.3.2.2 Modèle LSI (*Latent Semantic Indexing*)

Le modèle LSI (Deerwester *et al.*, 1990; Dumais, 1991), variante du modèle vectoriel, propose de transformer la représentation traditionnelle par mots-clés en une représentation plus « conceptuelle », plus « sémantique », qui vise à favoriser le rapprochement de documents et requêtes sémantiquement similaires. Partant du principe qu'une représentation vectorielle traditionnelle basée uniquement sur les mots contient trop de bruit (*i.e.* contient des termes non représentatifs du contenu textuel), il propose, en s'appuyant sur une décomposition en valeurs singulières de la matrice pondérée classique d'occurrences des termes d'indexation dans les documents de la collection, de créer un espace vectoriel plus petit où les dimensions ne sont plus représentées par les termes mais par une combinaison linéaire de ces termes. Ces combinaisons sont susceptibles de mieux faire ressortir les affinités sémantiques latentes entre les mots et, par conséquent, de mieux exprimer les concepts contenus dans les documents. L'utilisation du modèle LSI en RI consiste à traduire la requête de l'utilisateur dans ce nouvel espace. L'appariement d'un document et d'une requête revient alors à appliquer une mesure de similarité standard (*e.g.* la mesure du cosinus) entre les vecteurs dans l'espace réduit. Les documents peuvent comme dans le modèle vectoriel être classés selon leur pertinence par rapport à la requête.

Comme nous venons de le voir, un des avantages du modèle LSI est de permettre, par cette méthode de *clustering* de mots, une représentation plus sémantique des documents. En s'appuyant sur la décomposition en valeurs singulières de la matrice, elle permet d'obtenir un espace de représentation de dimension faible sans entraîner une perte trop importante d'information. Cette phase de réduction peut néanmoins s'avérer coûteuse en termes de calculs pour des matrices d'occurrences de grande dimension.

1.3.2.3 Modèle basé sur les réseaux de neurones

Une autre façon de modéliser en RI la relation entre les documents, la requête et les termes qu'ils contiennent est de s'appuyer sur le formalisme des réseaux de neurones (Baeza-Yates et Ribeiro-Neto, 1999). Un réseau de neurones en RI est généralement composé de plusieurs couches : une couche d'entrée qui désigne la requête (chaque neurone correspond à un de ses termes), une couche qui représente l'ensemble des termes de la collection (chaque neurone équivaut à un terme) et une couche documents où un nœud représente un document de la collection (*cf.* figure 1.4).

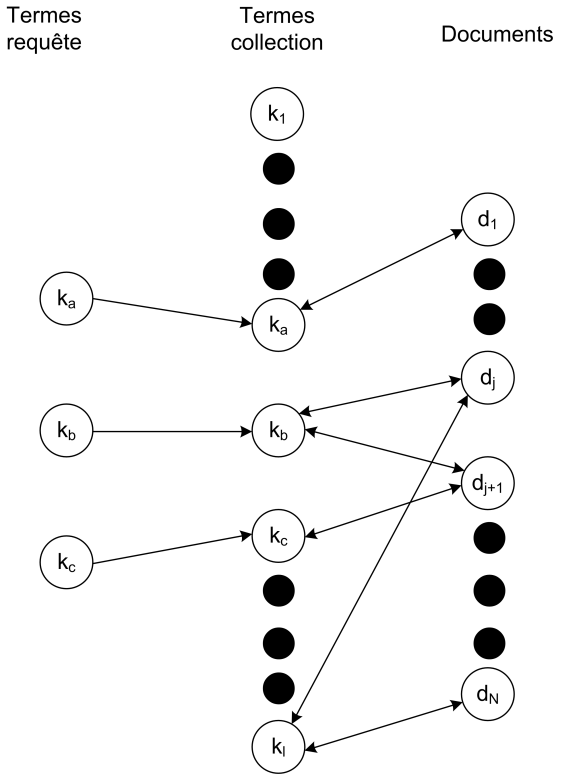

FIG. 1.4 – Modèle de réseau de neurones pour la RI (inspirée de (Baeza-Yates et Ribeiro-Neto, 1999))

À partir de cette représentation, le mécanisme d'appariement de la requête et des documents est relativement simple. L'activation initiale des neurones représentant la requête se propage vers les nœuds « termes » de la collection, qui, à leur tour, envoient des signaux aux nœuds documents, à travers les différentes connexions pondérées du réseau. Les documents qui ont reçu le plus de signaux sont considérés comme les plus pertinents. Un traitement itératif de propagation du signal peut venir compléter cette première phase d'activation : les nœuds des documents considérés comme pertinents génèrent alors de nouveaux signaux en direction des nœuds des termes de la collection (qui correspondent à des mots considérés comme très représentatifs du contenu de ces documents) qui envoient à leur tour de nouveaux signaux dirigés vers les nœuds des documents. Cette seconde phase, qui correspond à une forme de *relevance feedback* (Wilkinson *et al.*, 1992), permet de retrouver des documents dont les termes ne sont pas nécessairement contenus dans la requête.

Le principal intérêt des réseaux de neurones est qu'ils permettent de représenter diverses relations entre les termes (*e.g.* les relations de synonymie ou de voisinage (Boughanem, 1992; Boughanem *et al.*, 1998)), entre les documents (*e.g.* la similitude ou la référence) ou entre les termes et les documents (*e.g.* la fréquence, le poids...) (Sauvagnat, 2005). Leur fonctionnement en « boîte noire » constitue une de leurs principales limites puisque les résultats obtenus sont parfois difficilement interprétables par l'utilisateur, celui-ci ne comprenant pas toujours pourquoi tel document lui a été retourné.

Nous venons de présenter trois exemples de modèles algébriques utilisés en RI pour permettre aux systèmes d'établir un lien pertinent entre l'information recherchée et l'ensemble des documents disponibles. La relative simplicité des principes mis en œuvre (la correspondance entre une requête et un document est vue comme une mesure de similarité dans un espace vectoriel) et l'usage de techniques déjà bien établies sont les principaux atouts de ces modèles.

1.3.3 Modèles probabilistes

La particularité des modèles probabilistes, qui constituent avec le modèle vectoriel l'un des formalismes les plus utilisés aujourd'hui en RI, est de représenter la similarité d'un document vis-à-vis d'une requête par une probabilité de pertinence. Il s'agit plus précisément de répondre à la question suivante : étant donné un document d et une requête q, quelle est la probabilité que d soit pertinent pour q ? La principale différence de la modélisation probabiliste avec les modèles algébriques est qu'elle ne tient pas compte des autres documents pour évaluer la pertinence d'un texte en particulier. Chaque document est considéré individuellement par rapport au besoin d'information de l'utilisateur. Cette caractéristique (Rijsbergen, 1986) présente l'avantage d'éviter le problème de la haute dimensionnalité (*curve of dimensionality*) puisque lors de la recherche, tout se passe comme si la représentation des documents était réduite à un espace porté par les termes de la requête.

Après avoir présenté les modèles probabilistes classiques et les fondements théoriques sur lesquels ils s'appuient, nous proposons une brève description d'une modélisation fondée sur les réseaux bayésiens, puis terminons en évoquant une approche plus récente basée sur les modèles de langues.

1.3.3.1 Fondements des modèles probabilistes

Les modèles probabilistes définissent la notion de pertinence comme un critère binaire : un document est soit pertinent (noté désormais R pour *relevance*) soit non pertinent (NR). S'il est possible de déterminer les probabilités $P(R|d)$ (resp. $P(NR|d)$), *i.e.* la probabilité qu'un document d soit pertinent (R) (resp. non pertinent (NR)) (sous-entendu par rapport à une requête q), le système peut alors procéder au classement des documents selon ces deux probabilités. Cette idée repose sur le principe d'ordre des probabilités (*Probability Ranking Principle*) défini par Robertson (1977) qui affirme que la présentation des documents à l'utilisateur dans l'ordre décroissant des probabilités est optimale dans un cadre de RI. Selon ce principe, les documents retournés sont ceux dont la probabilité de pertinence est supérieure à la probabilité de non pertinence : $P(R|d) > P(NR|d)$. Pour pouvoir être directement calculables, les deux probabilités $P(R|d)$ et $P(NR|d)$ sont transformées selon le théorème de Bayes, et peuvent être réécrites de la manière suivante :

$$P(R|d) = \frac{P(d|R)P(R)}{P(d)} \text{ et } P(NR|d) = \frac{P(d|NR)P(NR)}{P(d)}$$

où $P(d|R)$ (resp. $P(d|NR)$) représente la probabilité que le document d fasse partie de l'ensemble des documents pertinents R (resp. des document non pertinents NR), $P(R)$ (resp. $P(NR)$) est la probabilité qu'un document choisi au hasard soit pertinent (resp. non pertinent), et $P(d)$ correspond à la probabilité qu'un document soit choisi. Après simplification, le calcul du score de correspondance entre un document et une requête peut être noté : $RSV(d,q) \approx \frac{P(d|R)}{P(d|NR)}$. Pour estimer la probabilité qu'un document soit pertinent, on s'appuie généralement sur la probabilité de pertinence de ses termes individuels : on cherche à calculer pour chaque terme t_i de la requête la probabilité qu'un document qui contienne t_i soit pertinent (resp. non pertinent), *i.e.* $P(t_{i=1}|R)$ (resp. $P(t_{i=1}|NR)$). Une des hypothèses fortes de ce modèle (qui est également une de ses limites) est qu'il suppose l'indépendance entre les termes (*i.e.* chaque terme qui apparaît dans un document est indépendant des autres termes). Sans entrer dans le détail des calculs (pour une explication plus précise, se référer à (Robertson et Spärck Jones, 1976)), le principal problème de ces modèles réside dans l'estimation des probabilités $P(t_{i=1}|R)$ et $P(t_{i=1}|NR)$.

Une solution consiste à utiliser un échantillon de documents où les hypothèses de pertinence sont déjà connues. R et NR correspondent alors à l'ensemble des documents pertinents et non pertinents dans cet échantillon. Le calcul de $P(t_{i=1}|R)$ (resp. $P(t_{i=1}|NR)$) est obtenu en appliquant une loi de distribution sur cet échantillon de documents. Des détails concernant l'estimation de ces probabilités et les fonctions de correspondance utilisées pour apparier les documents et la requête sont présentés dans (Spärck Jones

et al., 2000). D'autres mesures ont été proposées pour l'estimation des probabilités des termes sans recourir à des informations de pertinence *a priori* (Croft et Harper, 1979; Robertson et Walker, 1994). Ces méthodes sont généralement couplées à un traitement de *pseudo-relevance feedback* pour affiner les résultats obtenus initialement.

De nombreuses études ont cherché à apporter des améliorations à ce modèle de base. Nous pouvons citer les travaux de (Robertson et Walker, 1994) qui sont notamment à l'origine de la formule du $BM25$ (une approximation du modèle 2-Poisson) dont la principale caractéristique est d'intégrer dans la modélisation probabiliste la fréquence des termes dans les documents et requêtes et la longueur des documents. Cette mesure est à la base d'un des SRI les plus performants à l'heure actuelle : le système OKAPI (Robertson et Walker, 1994).

1.3.3.2 Réseaux bayésiens

Un autre formalisme probabiliste possible pour modéliser le processus de RI est celui des réseaux bayésiens. Deux types principaux de réseaux bayésiens ont ainsi été développés : les réseaux d'inférence (Turtle et Croft, 1991) (le système INQUERY (Callan *et al.*, 1992) est un exemple de SRI utilisant ce formalisme) et les réseaux de croyance (Ribeiro et Muntz, 1996). Nous détaillons ici plus précisément le premier type de modèle.

Un réseau d'inférence est représenté par un graphe acyclique orienté où les nœuds représentent les événements (ou variables) et les arcs, les relations (inférentielles) éventuellement pondérées entre ces nœuds. Appliqué à la RI, un réseau d'inférence (dont un

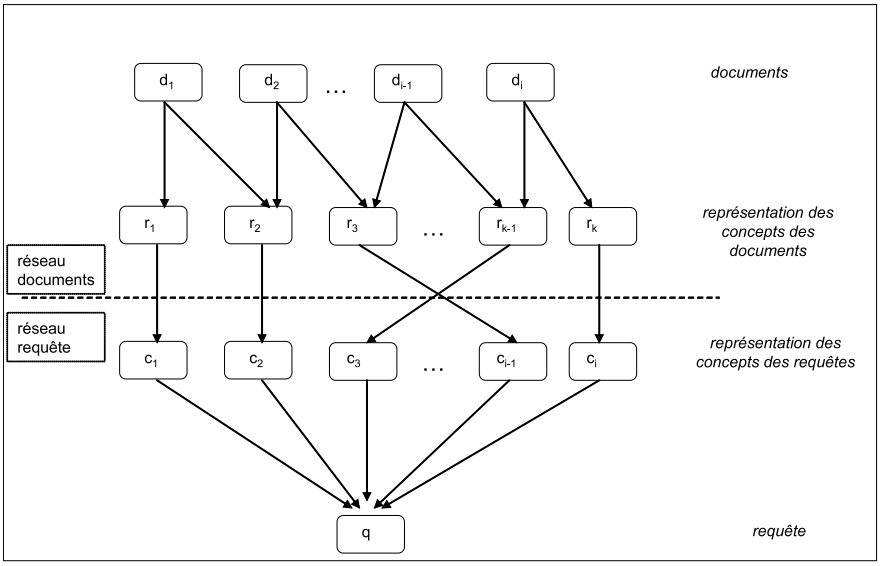

FIG. 1.5 – Exemple de réseau d'inférence pour la RI (Turtle et Croft, 1991)

exemple est donné en figure 1.5) se divise en deux sous-réseaux. Le sous-réseau docu-

ments se décompose en plusieurs niveaux hiérarchiques[5]. Le premier niveau représente les documents de la collection. Chaque nœud d_j correspond à la probabilité d'observer un document de la collection. Le second niveau correspond à la représentation des concepts, caractérisés le plus souvent par les termes des documents. Chaque nœud r_k désigne la probabilité qu'un concept (un terme) ait été observé étant donné l'ensemble de ses nœuds parents (notée $P(r_k|d_j)$). Cette probabilité prend en compte la fréquence des termes dans les documents, dans la collection ou d'autres formes de pondération. Le sous-réseau requête (reconstruit pour chaque nouvelle demande d'information) peut comporter également deux niveaux d'abstraction[6] : le niveau « requête » (noté q) formé de nœuds toujours vrais représentant la proposition qu'un besoin d'information de l'utilisateur soit satisfait ; le niveau « concept » dans lequel les nœuds c_i, vrais ou faux, représentent celle qu'un concept (terme) ait été observé ou non dans un document. Les deux sous-réseaux sont reliés par des arcs entre les nœuds concepts de la requête (c_i) et les nœuds concepts des documents (r_k). L'arc $r_k - c_i$ porte une valeur égale à la croyance qu'un nœud c_i ait été observé dans un document. Pour effectuer l'appariement entre une requête et un document, le réseau calcule, en activant un document à la fois et en propageant les valeurs de croyance de nœud en nœud, la probabilité que le document considéré réponde au besoin d'information de l'utilisateur. Les documents sont ensuite ordonnés selon cette probabilité.

Les principaux avantages des réseaux d'inférence est de permettre d'une part la combinaison de multiples représentations d'un même document ou d'un même besoin d'information de l'utilisateur — offrant par exemple la possibilité d'intégrer plusieurs formulations différentes d'une même requête — et, d'autre part, d'implémenter différentes stratégies de recherche en parallèle (du fait notamment de la séparation du réseau requête et du réseau documents) (*cf.* par exemple (Rajashekar et Croft, 1995)). Leur principal inconvénient réside dans le calcul des probabilités, nécessitant un temps exponentiel par rapport au nombre de termes de la requête. Les réseaux de croyance proposés par (Ribeiro et Muntz, 1996) sont une généralisation des réseaux d'inférence et s'en distinguent principalement par l'orientation inverse donnée au sens des arcs du réseau (les valeurs se propagent de la requête vers les documents).

1.3.3.3 Modèles de langue

L'utilisation en RI des modèles de langue, basés également sur une approche probabiliste, est récente (Ponte et Croft, 1998). Les modèles de langue, qui calculent sur des corpus d'apprentissage des probabilités de succession de mots, sont habituellement utilisés en reconnaissance de la parole et en traduction (Manning et Schütze, 2000). En RI, leur principe est le suivant :

[5] Nous considérons ici seulement deux niveaux. Les documents au sein des réseaux d'inférence peuvent cependant être décrits selon d'autres niveaux d'abstraction (Turtle et Croft, 1991).

[6] Les réseaux d'inférence offrent néanmoins la possibilité d'ajouter un niveau intermédiaire qui permet d'exprimer le besoin d'information de l'utilisateur à l'aide de plusieurs requêtes plus ou moins disjointes, utilisées par exemple pour modéliser ce besoin sous forme d'un profil d'intérêt.

- un document est considéré comme un échantillon d'un langage particulier. Un modèle de langue (noté M_d) est donc entraîné pour chaque document de la collection ;
- la requête (notée q) est vue comme un processus de génération d'une phrase dans les différents langages des documents (*i.e.* leur modèle de langue) ;
- la fonction de correspondance entre une requête et un document est définie en estimant la probabilité que la requête q puisse être générée par le modèle de langue d'un document M_d ;
- les documents retournés à l'utilisateur sont alors classés dans l'ordre décroissant de la probabilité $P(q|M_d)$.

La particularité de ces modèles, par rapport aux approches probabilistes classiques, est qu'ils ne cherchent pas à modéliser la notion de pertinence. Les modèles probabilistes exposés précédemment tentent en effet d'évaluer la probabilité qu'un document soit pertinent étant donnée une requête. Dans les modèles de langue, la pertinence d'un document vis-à-vis d'une requête est considérée uniquement comme la probabilité que cette dernière puisse être générée par le modèle de langue d'un document.

La probabilité qu'un mot q d'une requête soit généré par le modèle de langue d'un document d est estimée par comptage, et revient généralement (dans le cas d'un modèle unigramme) à calculer la fréquence des termes de q dans d. Pour éviter d'assigner une probabilité nulle à un document qui ne contiendrait aucun terme de la requête, on utilise des techniques de lissage (*smoothing*). Leur principe est de redistribuer une partie de la totalité de la masse des n-grammes vus dans le corpus d'entraînement aux n-grammes non vus, permettant à ces derniers de recevoir ainsi une probabilité non nulle. Les travaux de (Alvarez *et al.*, 2003) passent en revue les différentes techniques de lissage généralement utilisées en RI, qui influent fortement sur les performances des SRI.

Plusieurs raisons conduisent à penser que ces modèles, basés sur un cadre mathématique simple, sont très prometteurs. Ils présentent notamment l'avantage, contrairement aux modèles algébriques, d'intégrer au sein d'un seul modèle la phase d'indexation et de recherche, réduisant ainsi les calculs. Ils introduisent également une nouvelle fonction de classement basée sur la génération de la requête. De plus, ils ne nécessitent, à la différence des modèles probabilistes, aucun jugement de pertinence pour fonctionner. D'autre part, les recherche actuelles tendent à s'orienter vers la conception de modèles de langue susceptibles de dépasser l'hypothèse d'indépendance entre les termes. Plusieurs propositions ont en effet été faites pour intégrer des relations de dépendances entre les termes (Song et Croft, 1999; Miller *et al.*, 1999; Alvarez *et al.*, 2003; Srikanth et Srihari, 2002; Maisonnasse, 2005). La difficulté d'incorporer les stratégies de *relevance feedback* et les préférences de l'utilisateur constitue toutefois l'un des points faibles de ces modèles.

Nous venons de présenter quelques-uns des principaux modèles intégrés au sein des SRI. Comme nous l'avons déjà dit, cette description n'est pas exhaustive : nous aurions pu également évoquer les modèles logiques (Rijsbergen, 1986; Nie, 1990; Chevallet, 2004) par exemple. Nous nous sommes contentée de décrire les principes généraux de fonctionnement de ces modèles, sans spécifier plus précisément les mécanismes mis en

œuvre pour améliorer les fonctionnalités de base. Cette présentation a néanmoins tenté de mettre en avant les principaux points forts et points faibles des modèles de RI. L'une des limites communes à la plupart d'entre eux réside dans le fait que l'ordre séquentiel des mots des textes (et des questions) est généralement ignoré, les documents et requêtes étant globalement représentés comme des « sacs de mots », ce qui conduit les SRI à ne pas tenir compte des diverses relations que les termes peuvent entretenir les uns avec les autres. Une autre limite de ces modèles est liée à la définition même de la notion de pertinence qu'ils tentent de modéliser. Les fonctions de correspondance utilisées pour parvenir à déterminer de la pertinence ou de la non pertinence d'un document s'appuient essentiellement sur la présence ou l'absence des termes de la requête dans les documents. Or un tel mécanisme est nécessairement limité d'une part parce qu'il n'est pas évident qu'un document qui ne contient aucun des termes de la requête soit forcément inintéressant pour l'utilisateur (il est en effet possible d'exprimer une même idée de manières différentes) et, d'autre part, parce que la présence des termes de la requête dans un texte ne suffit pas à attester de sa pertinence par rapport à l'information recherchée par l'utilisateur (notamment à cause des problèmes d'ambiguïté des mots de la langue). Toutes ces faiblesses entraînent des performances limitées pour les SRI, évaluées à l'aide de mesures et de protocoles rigoureux (estimation de la pertinence « système »). Ces techniques d'évaluation font l'objet de la section suivante.

1.4 Techniques d'évaluation des performances des SRI

Comme nous l'avons déjà évoqué, la pertinence est une notion très complexe à évaluer. En effet, elle dépend fortement de l'utilisateur, qui est véritablement le seul à savoir si le document retourné par le système correspond à son besoin d'information initial. Il est cependant indispensable de disposer de techniques d'évaluation solides qui, en définissant des mesures précises, permettent de juger de la performance des SRI, quels que soient les méthodes d'indexation, de recherche ou les modèles qu'ils implémentent. Pour cela, ces techniques s'appuient essentiellement sur l'estimation de la qualité des informations retrouvées par les systèmes, *i.e.* les documents retrouvés sont-ils pertinents ou non pertinents ? D'autres critères peuvent toutefois être pris en considération, comme par exemple :

- le temps mis par le système pour fournir des réponses à l'utilisateur,
- l'effort effectué par l'utilisateur pour obtenir l'information recherchée (*e.g.* le nombre de requêtes qu'il a dû formuler avant d'avoir le résultat recherché),
- la qualité de la présentation des résultats par le système (*e.g.* à partir de la liste de résultats fournis par le système, combien de documents l'utilisateur a-t-il dû parcourir avant de trouver le document recherché ?).

Nous évoquons dans cette partie uniquement les techniques liées à l'évaluation de la qualité des informations retrouvées par le SRI. De nombreux travaux s'intéressent également à la possibilité d'intégrer l'utilisateur dans ces mesures d'évaluation (*cf.* notamment (Korhage, 1997)).

Pour pouvoir être évalués et comparés les uns aux autres, les SRI doivent être appliqués sur un même jeu de données et utiliser des méthodologies d'évaluation identiques. De nombreux projets d'évaluation ont vu le jour depuis le début des années 70 (*e.g.* le projet Cranfield) afin de construire des collections de test complètes et définir des protocoles d'expérimentations précis. Aujourd'hui, l'initiative la plus importante est la campagne d'évaluation TREC (*Text REtrieval Conference*) qui, en plus de fournir des collections de test volumineuses, propose une infrastructure bien définie pour l'évaluation des méthodologies de recherche.

Nous proposons donc dans un premier temps de revenir plus en détail sur les spécificités de la campagne TREC et, plus précisément, sur les collections de test qu'elle offre et que nous avons utilisées pour nos travaux. Dans un second temps, nous présentons les diverses mesures d'évaluation traditionnellement appliquées en RI pour estimer les capacités des SRI à retrouver les documents pertinents recherchés par l'utilisateur, dont nous nous servons dans les chapitres suivants pour évaluer la qualité des techniques que nous proposons.

1.4.1 Campagne d'évaluation TREC

L'objectif des campagnes d'évaluation TREC est de proposer une plate-forme qui réunit des collections de test, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche (Boughanem, 2003) pour mesurer les performances des SRI et, plus précisément, les méthodes d'indexation et de recherche et les modèles de RI sous-jacents. Parmi les différentes tâches de recherche proposées par TREC pour l'évaluation de problèmes spécifiques en RI[7], on peut citer par exemple, outre la tâche de RI « standard », les tâches de question-réponse (les systèmes doivent retourner à l'utilisateur non plus des documents susceptibles de contenir la réponse mais la réponse à une question précise), de RI translinguistique (*i.e.* retrouver des documents dans une langue différente de celle de la requête), de recherche appliquée à de grands corpus (25 giga-octets), à la vidéo...

La particularité de TREC est de mettre à disposition des participants des collections de test volumineuses, complétées et modifiées d'une campagne à une autre (à raison d'une par an). Ces collections de données sont composées essentiellement d'un ensemble important de documents, de *topics* correspondant aux besoins d'information de l'utilisateur, et d'un ensemble de jugements de pertinence, trois éléments que nous présentons brièvement ci-dessous.

1.4.1.1 Collections de documents

Dans le cadre d'une tâche de RI « classique », les documents proposés sont en texte intégral et faiblement structurés (seule une balise titre et éventuellement des marques de paragraphes sont indiquées). Ils ont été volontairement choisis pour leur variété thématique et sont généralement issus (pour plus de la moitié d'entre eux) d'articles de

[7] Pour plus de détails, se référer à l'adresse suivante : `http://trec.nist.gov/tracks.html`.

journaux (*Financial Times*, *Los Angeles Times*, *Wall Street Journal*...) ou de collections de documents légaux, de brevets ou de documents spécialisés dans un domaine particulier (*e.g.* l'informatique). Leur longueur est également variable. Selon (Bommier-Pincemin, 1999), la taille moyenne des documents (pour la majorité d'entre eux) est d'environ 300 mots. Le volume de ces collections est variable d'une année à une autre. Néanmoins, à titre d'exemple, la collection utilisée à TREC-8 comprenait environ 530 000 documents (soit 1904 MB).

1.4.1.2 *Topics*

Un *topic* est considéré dans ces collections comme une description enrichie du besoin d'information de l'utilisateur. Il se décompose en plusieurs éléments (dont le nombre varie selon les campagnes). On distingue le plus souvent le champ *titre* (deux ou trois mots-clés représentatifs du contenu du *topic*), le champ *description* qui correspond à la requête proprement dite (généralement représenté par deux ou trois phrases) et le champ *narrative* qui détaille précisément le type de document attendu en réponse. Généralement, seuls les éléments *titre* et *description* sont utilisés lors des expérimentations. Afin de ne pas biaiser les résultats, un jeu de 50 requêtes est le plus souvent considéré lors de l'évaluation des systèmes. Dans ce mémoire, les termes *topic* (regroupant donc les champs *titre* et *description*) et *requête* seront employés sans distinction.

1.4.1.3 Jugements de pertinence

Le principal atout de ces collections de test est de proposer des jugements de pertinence associés aux documents et requêtes. Un jugement de pertinence est « l'assignation d'une valeur de pertinence par un juge à un moment défini » (Mizarro, 1997). Autrement dit, à chaque requête est associé un certain nombre de documents qui ont été jugés au préalable comme pertinents. C'est à partir de cette « vérité terrain » que l'efficacité des systèmes va pouvoir être évaluée. Le nombre de documents disponibles étant trop conséquent, il est impossible de demander à un humain (même à plusieurs) d'évaluer la pertinence de chaque texte étant donnée une requête particulière. Pour construire ces indicateurs de pertinence, c'est une technique de jugement *a posteriori* qui est utilisée, nommée « pooling » (*jugement sur échantillon*) (Spärck Jones et Rijsbergen, 1975). Elle consiste tout d'abord à lancer en parallèle les SRI participant à la campagne d'évaluation, puis à retenir l'union des 100 premières réponses fournies par chaque SRI, et cela pour chaque requête. Des experts analysent ensuite l'ensemble de ces réponses puis assignent une valeur de pertinence binaire (1 si le document est pertinent, 0 sinon) à chaque document. Bien que cette technique d'échantillonnage soit largement discutable (Buckley et Voorhees, 2000; Zobel, 1998), il a été démontré qu'elle était viable et valide pour évaluer les performances de SRI.

Les collections de test et les méthodes d'évaluation proposées dans la campagne TREC font l'objet de nombreuses critiques. Elles proposent en effet uniquement des mesures quantitatives au détriment de mesures qualitatives prenant en compte la sa-

tisfaction de l'utilisateur. Elles utilisent également des données assez éloignées des recherches effectuées en réalité par un utilisateur *lambda*. Malgré cela, TREC reste sans conteste la référence en matière d'évaluation de SRI. De plus, son existence a permis d'encourager le développement d'autres campagnes d'évaluation, telles que CLEF (*Cross Language Evaluation Forum*) pour l'évaluation de SRI multilingues, Amaryllis (pour la langue française) ou INEX (*Initiative for the Evaluation of XML Retrieval*) pour la RI structurée.

1.4.2 Mesures d'évaluation de SRI

Comme nous l'avons évoqué précédemment, l'évaluation d'un SRI consiste à estimer son efficacité à retrouver des documents pertinents. Diverses mesures ont été proposées pour répondre à cet objectif. Elles s'appuient sur plusieurs hypothèses fortes (Piwowarski, 2003; Claveau, 2003) qui doivent être prises en considération lors de leur utilisation, parmi lesquelles :

- la *pertinence binaire* : un document est considéré soit pertinent, soit non pertinent. Aucune graduation de pertinence n'est prise en compte dans la tâche d'évaluation ;
- les *jugements de pertinence absolus* : aucune remise en question de ces jugements n'est possible ;
- la *non-additivité* : chaque document est évalué indépendamment des autres membres de la collection. La pertinence d'un document ne dépend pas des autres textes répondant à la requête ;
- l'*absence de mémoire* : un document reste pertinent même si un autre au contenu similaire a déjà été présenté à l'utilisateur.

Bien que certaines de ces hypothèses soient contestables, elles sont toutes considérées comme avérées dans les mesures d'évaluation que nous décrivons à présent.

1.4.2.1 Rappel et précision

Les deux mesures communément utilisées depuis plus de 30 ans pour évaluer un système de recherche d'information sont le taux de *précision* et celui de *rappel* (Salton, 1992). Ces deux mesures peuvent être définies par :

$$\textit{précision} = \frac{\textit{nombre total de documents pertinents retrouvés par le système}}{\textit{nombre total de documents retrouvés par le système}}$$

$$\textit{rappel} = \frac{\textit{nombre total de documents pertinents retrouvés par le système}}{\textit{nombre total de documents pertinents dans la collection}}$$

D'une manière duale, on peut aussi définir les notions de *bruit* et de *silence* qui sont respectivement complémentaires de la précision et du rappel : $\textit{bruit} = 1 - \textit{précision}$ et $\textit{silence} = 1 - \textit{rappel}$. Un système idéal aurait un taux de rappel et de précision de 100%. Cela signifie qu'il devrait pouvoir retrouver tous les documents pertinents de la collection (rappel de 100%) et seulement ceux-ci (précision de 100%). En pratique (et comme les résultats des tests de Cranfield l'ont démontré (Cleverdon, 1967)), ces deux

mesures sont liées et varient de manière inversement proportionnelle. Il est donc peu significatif de mesurer la précision d'un système sans calculer son rappel et inversement.

Dans la plupart des modèles, ce sont des scores qui sont assignés aux documents retrouvés par le système et non une décision de pertinence binaire (le document est pertinent ou non). Pour en déduire une réponse binaire, il faut donc déterminer un seuil de score tel que les documents obtenant un score supérieur à ce seuil sont considérés comme pertinents, et ceux possédant un score inférieur comme non pertinents. Pour fixer ce seuil, on s'appuie généralement sur le nombre de documents qui doivent être considérés comme pertinents. Ce nombre est appelé DCV (pour *document cut off value*). Il est alors possible de calculer la précision et le rappel correspondant à un DCV donné — les valeurs les plus utilisées sont 5, 10, 15, 20, 30, 50, 100, 500... Si le DCV est fixé à 10 par exemple, la précision (notée P(10)) correspond au nombre de documents pertinents retrouvés parmi les 10 premiers ramenés par le système, le rappel (noté R(10)) est le nombre de documents pertinents retrouvés parmi les 10 premiers divisé par le nombre total de documents pertinents de la collection. Les performances d'un système peuvent alors être représentées par une courbe rappel/précision comme celle présentée en figure 1.6.

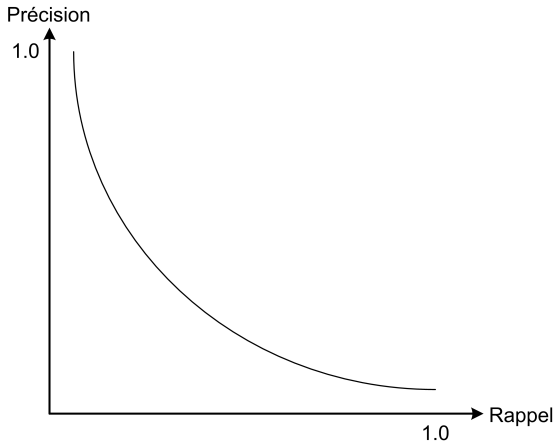

FIG. 1.6 – Exemple de courbe rappel/précision

Il est également fréquent, lorsque les valeurs exactes de rappel ne peuvent être atteintes, d'appliquer une interpolation sur ces courbes. Elle consiste à lisser la courbe initiale pour qu'elle soit décroissante : la valeur interpolée de la précision pour un point de rappel i est la précision maximale obtenue pour un point supérieur ou égal à 1. Le principal avantage est de définir la précision sur des valeurs standardisées.

Le calcul de la précision et du rappel s'effectue pour chaque élément de la liste des documents retrouvés par le SRI. Pour évaluer l'efficacité globale d'un système, il peut être utile de disposer de mesures uniques qui synthétisent en une seule grandeur la performance du SRI. Dans ce but, plusieurs mesures ont été proposées.

1.4.2.2 Mesures complémentaires

La *précision moyenne interpolée* (notée IAP pour *Interpolated Average Precision*) est une mesure décrivant la précision globale du système évalué sur une requête. Elle consiste à calculer la précision des résultats sur onze points, correspondant aux DCV pour lesquels le rappel vaut 0, 10, 20, 30, 40, 50, 60, 70, 80, 90 et 100%. Si ces points ne sont pas atteints en fixant un DCV, les mesures sont alors interpolées. La moyenne de ces 11 précisions forme la précision moyenne interpolée.

La *précision moyenne non interpolée* (représentée généralement par différents acronymes (MAP, NIAP, ou UAP pour *Uninterpolated Average Precision*)) représente la moyenne des précisions non interpolées obtenues pour tous les DCV correspondant au rang d'un bon document dans la liste des réponses. Si un document pertinent se trouve en position (rang) 10, la valeur qui lui est associée pour le calcul de la précision moyenne correspond au taux de documents pertinents parmi les 10 premiers documents. La MAP est la moyenne arithmétique de toutes ces valeurs.

La *R-précision* mesure, pour une requête, la précision obtenue pour un nombre donné de documents retrouvés par le système. Ce nombre est fixé pour chaque requête en fonction du nombre de documents pertinents dans la collection. La R-précision est intéressante lorsque la collection comporte un nombre important de documents pertinents (comme c'est le cas dans les collections de TREC par exemple).

La *F-mesure* (Rijsbergen, 1979) est définie comme la moyenne harmonique pondérée du taux de rappel et du taux de précision :

$$F\alpha(s) = \frac{P(s)R(s)}{(1-\alpha)P(s) + \alpha R(s)},\ 0 \leq \alpha \leq 1.$$

où $P(s)$ et $R(s)$ représentent respectivement les résultats de précision et de rappel en fonction d'un seuil s (DCV), et α est la valeur qui traduit l'importance que l'on souhaite donner aux deux taux. Dans le cas où un poids égal est attribué à la précision et au rappel, α vaut $0,5$ et la F-mesure s'écrit :

$$F(s) = \frac{2P(s)R(s)}{P(s) + R(s)}$$

.

Les mesures présentées ici sont celles les plus fréquemment utilisées pour évaluer les performances des SRI. Une étude plus complète est proposée par (Mizarro, 1997). Ces différentes mesures font souvent l'objet de critiques liées principalement aux hypothèses évoquées en introduction de cette section. L'un des reproches concerne l'absence de prise en compte de l'utilisateur. Bien que plusieurs travaux visent à développer des mesures d'évaluation orientées utilisateur (Korhage, 1997) (en considérant par exemple le nombre de documents consultés par l'utilisateur avant de trouver le document recherché, ou le nombre de documents qui étaient jusqu'alors inconnus pour l'utilisateur (degré de nouveauté des documents)), ils restent en marge des méthodologies d'évaluation traditionnelles.

Enfin, il est souvent utile, lorsque l'on souhaite comparer deux SRI — le premier servant généralement d'étalon (*baseline*), le second étant le SRI à évaluer — de s'assurer

que les améliorations ou les dégradations constatées (pour le SRI à évaluer) ne sont pas le fait du hasard. Plusieurs tests statistiques sont donc employés en RI afin de vérifier que chacune des hausses ou des baisses observées lors des expérimentations est statistiquement significative. Parmi les plus fréquents, on trouve principalement le *paired t-test* (un dérivé du *t-test de Student*), ou le test non paramétrique de Wilcoxon (*paired Wilcoxon test*) jugé plus robuste et souvent préféré (Rijsbergen, 1979). Ces mesures, non détaillées ici[8], seront utilisées par la suite pour évaluer nos résultats.

1.5 Bilan : vers une RI plus linguistique

L'objectif de ce chapitre était d'introduire le domaine de la recherche d'information, de décrire plus particulièrement les principales étapes — l'indexation et la recherche — par lesquelles les systèmes doivent passer pour arriver à mettre en correspondance l'information recherchée par l'utilisateur et les documents de la base, de présenter les principaux modèles sur lesquels s'appuient les SRI, et de décrire les méthodes d'évaluation adoptées pour attester de la validité de tous les mécanismes implémentés au cœur de ces systèmes. Ces différents points ont permis de mettre en avant le problème central de la RI qui s'articule essentiellement autour de la pertinence, notion qui s'avère complexe tant par sa définition que par sa modélisation ou son évaluation. La plupart des travaux de recherche de RI s'orientent donc vers le même objectif : s'attacher à rapprocher la pertinence « système » de la pertinence « utilisateur », et cela, quelles que soient les voies empruntées pour y parvenir. Les performances des SRI participant aux campagnes d'évaluation TREC attestent cependant de la difficulté à parvenir à un tel objectif. Les taux de précision et de rappel obtenus par les meilleurs d'entre eux n'excèdent pas en effet les 30-40%, ce qui signifie entre autres que la majorité des documents retrouvés ne correspondent pas à la requête et que, par conséquent, les techniques utilisées ne sont pas suffisamment efficaces pour permettre de récupérer les documents pertinents de la collection.

En présentant dans ce chapitre les mécanismes de RI à la base de la plupart de ces systèmes, nous avons pu faire ressortir plusieurs limites pouvant expliquer ces résultats. La première est liée aux méthodes utilisées pour représenter les contenus textuels. Le passage du document (ou de la requête) en texte intégral à une représentation en « sac de mots », telle qu'elle est proposée par la plupart des modèles de RI, entraîne des pertes d'informations conséquentes. Cette représentation ne permet d'une part pas de prendre en compte l'ordre des mots puisque les documents sont déséquentialisés. Elle ignore d'autre part les relations susceptibles d'exister entre les mots, la plupart des modèles de RI s'appuyant sur l'hypothèse d'indépendance entre les termes. Or, dans la langue, les mots ne sont pas agencés les uns à la suite des autres par hasard ; chaque phrase a une structure significative ; chaque mot s'inscrit dans un contexte, entretient des relations avec les autres termes... Tous ces éléments participent fortement à la construction du sens, et leur absence de prise en compte a nécessairement des répercussions importantes sur les performances des systèmes.

[8] Pour plus de détails, se référer à (Hull, 1993; Claveau, 2003).

La seconde limite des mécanismes traditionnels de RI, fortement liée à la précédente, concerne les méthodes utilisées pour mettre en correspondance les documents et les requêtes, représentés par leurs « sacs de mots » respectifs. Comme nous l'avons évoqué, le processus d'appariement tel qu'il est actuellement utilisé par la plupart des systèmes peut se résumer à une simple comparaison de mots (*i.e.* des mots (éventuellement pondérés) des documents et ceux des requêtes). Les faiblesses de ce mécanisme de mise en correspondance résident principalement autour de deux problèmes. Le premier concerne la possibilité offerte par le langage naturel de formuler de différentes manières une même idée, un même concept. La principale conséquence d'un tel phénomène est qu'un document pourtant intéressant pour l'utilisateur ne contient pas nécessairement les mêmes mots que ceux qu'il a employés pour exprimer son besoin d'information. Un document pertinent peut ainsi contenir des termes « sémantiquement » proches de ceux de la requête mais toutefois différents (*e.g.* des synonymes : *voiture vs automobile*, des hyperonymes : *canari vs oiseau*, ou encore des termes ayant une forme morphologique différente : *indexer vs indexation*). Ce phénomène est à l'origine de la baisse du rappel des systèmes qui ne peuvent proposer à l'utilisateur certains documents pourtant pertinents. Le second problème, dual du premier, est celui de la polysémie des mots (*e.g.* le terme *orange* qui désigne à la fois la ville, le fruit, la couleur ou l'opérateur de téléphonie). L'ambiguïté qui en découle conduit à une baisse de précision des systèmes puisqu'elle entraîne potentiellement la récupération de documents non pertinents.

Compte tenu de ces difficultés liées principalement à la complexité du langage naturel, une solution souvent évoquée est de recourir à des unités plus fines que les chaînes graphiques pour représenter les documents et requêtes. Ces nouvelles unités sont obtenues par une analyse linguistique des contenus textuels effectuée par le biais de techniques du traitement automatique des langues (TAL) et doivent permettre aux SRI, en proposant des descriptions plus riches des contenus textuels, une mise en correspondance plus pertinente des documents et requêtes.

Le chapitre suivant propose une synthèse des principaux apports des techniques du TAL à la RI à travers un tour d'horizon des diverses tentatives de couplage de ces deux domaines déjà réalisées.

Chapitre 2

Apport de techniques du TAL en RI

Résumé : Les techniques issues du traitement automatique des langues permettent d'extraire des textes des informations plus riches que de simples mots-clés. Ces connaissances de nature morphologique, syntaxique et sémantique ont été partiellement exploitées en RI pour améliorer les mécanismes traditionnels de représentation des contenus des documents et requêtes, les méthodes d'appariement et le processus de recherche. Ce chapitre dresse un bilan de l'impact de ces différentes informations linguistiques obtenues par des techniques du TAL sur les systèmes de recherche d'information et sur leurs performances, à travers un tour d'horizon des nombreux travaux qui se sont intéressés à ce couplage TAL et RI.

Mots-clés : traitement automatique des langues, recherche d'information, informations morphologiques, syntaxiques et sémantiques, indexation, extension de requêtes.

2.1 Introduction

Les mécanismes classiques de RI, nous l'avons vu, se heurtent essentiellement à des difficultés de nature linguistique. Les « sacs de mots » utilisés pour représenter les documents et les requêtes durant tout le processus de recherche d'information ne véhiculent tout d'abord que partiellement leur contenu sémantique. En effet, en ne prenant en compte ni les relations et dépendances qu'entretiennent les termes les uns avec autres, ni l'ordre des mots, ils engendrent des pertes d'informations qui sont en grande partie responsables des performances limitées des SRI. La mise en correspondance entre l'information recherchée par l'utilisateur et l'ensemble des documents disponibles ne peut pas non plus être réduite, comme le proposent actuellement la plupart des modèles de RI, à une simple comparaison de chaînes de caractères. La richesse et la complexité de la langue exigent un mécanisme d'appariement plus fin, capable de considérer comme potentiellement pertinent un document qui ne contiendrait aucun des termes utilisés dans la requête et, réciproquement, de rejeter un document effectivement non intéressant même si ce dernier possède les chaînes contenues dans la requête.

Pour pallier ces limites, principalement articulées autour de la notion de mot, une solution assez naturelle est de recourir à des unités textuelles plus riches, susceptibles de mieux représenter les documents et requêtes et de permettre par conséquent aux SRI une meilleure préhension des contenus. Ces nouvelles unités informationnelles peuvent être obtenues à partir d'une analyse linguistique des documents et requêtes, effectuée par le biais de techniques du traitement automatique des langues. Cette analyse a pour principal avantage de ne plus considérer les mots comme de simples graphies mais comme des entités linguistiques à part entière, *i.e.* comme des unités susceptibles de posséder plusieurs sens, de subir des variations de formes, de structures, d'entretenir des relations avec d'autres unités... L'exploitation de ces nouvelles unités par les SRI devrait donc permettre une amélioration de leurs performances.

Les traitements linguistiques peuvent intervenir à différents niveaux dans un SRI. Ils contribuent d'une part, en exploitant des connaissances linguistiques extraites des textes, à améliorer le processus d'indexation des documents et requêtes, et à créer une représentation plus riche de leur contenu ; cette représentation vise à obtenir une mise en correspondance plus fine et donc plus pertinente que de simples mots-clés entre l'information recherchée par l'utilisateur et les documents de la collection. Ils ont, d'autre part, pour objectif d'améliorer le processus de recherche des SRI en enrichissant les requêtes par des informations complémentaires. Ces informations permettront aux systèmes de mieux cerner le besoin de l'utilisateur et de retrouver par conséquent davantage de documents intéressants.

Le rapprochement des communautés TAL et RI a donné lieu à un nombre important de travaux de recherche. Nous proposons dans ce chapitre de présenter une synthèse de ces contributions, à travers un tour d'horizon des diverses tentatives de couplage TAL et RI qui ont déjà été réalisées[1]. En TAL, les informations linguistiques extraites des textes peuvent être de natures différentes. On distingue généralement trois principaux niveaux d'analyse linguistique : les niveaux morphologique, syntaxique et sémantique. L'analyse morphologique s'intéresse à la forme des mots ; l'analyse syntaxique étudie la façon dont les mots se combinent pour former des syntagmes et des phrases ; et l'analyse sémantique a pour objet l'étude du sens véhiculé par les mots. Afin de dresser un état de l'art aussi exhaustif que possible des différents apports du TAL à la RI, nous choisissons de reprendre ce découpage et de nous attarder plus précisément sur la façon dont les informations d'ordre morphologique (section 2.2), syntaxique (section 2.3) et sémantique (section 2.4) extraites (entre autres) des documents et requêtes ont été jusqu'à présent intégrées au sein des SRI. Nous proposons enfin (section 2.5) un bilan des principaux résultats d'exploitation des connaissances linguistiques en RI qui se dégagent de cet état de l'art, bilan sur lequel nous nous appuyons pour établir nos propres pistes de recherche.

[1] Les travaux couplant le TAL à la RI ne sont présentés ici que dans les grandes lignes. Les diverses expérimentations proposées font l'objet d'une description plus détaillée dans notre rapport de recherche (Moreau et Sébillot, 2005).

2.2 Apport de connaissances morphologiques en RI

Le principal intérêt en RI d'une « analyse » morphologique appliquée aux documents et requêtes est de pouvoir reconnaître que les mots *produire*, *produit*, *producteur* et *productrice*, bien que graphiquement différents, sont en réalité différentes formes d'un même mot — on parle de variantes morphologiques. L'appariement de ces formes, sémantiquement proches, peut par conséquent s'avérer pertinent.

Cette section s'intéresse donc au traitement de la variation morphologique en RI. Après un bref rappel de quelques notions fondamentales en morphologie (section 2.2.1), nous présentons les différentes expériences qui ont exploité des connaissances morphologiques au sein de systèmes (section 2.2.2), en découpant notre analyse selon le type d'outils utilisés pour mettre en évidence ces informations à partir des textes. Nous proposons enfin (section 2.2.3) un bilan de leur prise en compte en RI.

2.2.1 Quelques notions utiles de morphologie

La morphologie concerne la structure des mots (Polguère, 2003), c'est-à-dire les combinaisons de morphèmes (*i.e.* les plus petites unités de sens) qui les forment. Ces morphèmes sont soit lexicaux, soit grammaticaux (les affixes), ces derniers ne pouvant être utilisés seuls et se combinant par conséquent aux précédents. La morphologie peut généralement être vue sous trois angles différents. La morphologie flexionnelle, tout d'abord, étudie les flexions des mots, *i.e.* le passage d'un mot à un autre par ajout ou suppression des marques singulier/pluriel (*chien* et *chiens*), masculin/féminin (*chien* et *chienne*) pour les noms et adjectifs, des marques de mode/temps/personne pour les verbes ou de cas pour les langues à déclinaisons (allemand, polonais, russe...). La morphologie dérivationnelle s'intéresse à la formation de nouvelles unités lexicales à partir de morphèmes lexicaux et d'affixes (*i.e.* préfixes, suffixes et infixes pour certaines langues) dits dérivationnels. Elle explique par exemple la formation de *reconstruction* à partir du morphème lexical *construct*, du préfixe *re-* et du suffixe *-ion*. La morphologie compositionnelle, qui ne sera pas abordée ici, porte également sur la formation de nouvelles unités, mais par juxtaposition de plusieurs morphèmes lexicaux (*e.g.* dans *portefeuille*).

Un mot pouvant donc avoir plusieurs formes de sens proches (*e.g.* *transformer*, *transforme*, *transformateur*, *transformation*...), il s'avère parfois pertinent, et particulièrement dans un cadre de RI, de pouvoir les rassembler autour d'une forme unique. Le rôle de l'analyse morphologique est de déterminer les diverses formes effectives d'un même mot, ou sa famille morphologique[2] — par exemple en l'extrayant automatiquement de documents —, et de procéder à leur normalisation, c'est-à-dire à un recodage des diverses variantes du mot par une forme unique. Si l'on s'oriente vers la morphologie flexionnelle, la forme unique sera le lemme, *i.e.* une forme unique débarrassée de ses flexions ; le traitement associé est la lemmatisation. Le recours à la morphologie dérivationnelle aura pour principal résultat la normalisation des formes autour d'une racine[3] ou d'un

[2] Une famille morphologique peut être définie comme regroupant un ensemble de mots de racine morphologique commune, obtenue à partir d'une analyse flexionnelle et dérivationnelle.

[3] « On appelle racine l'élément de base, irréductible, commun à tous les représentants d'une même famille de mots à l'intérieur d'une langue ou d'une famille de langue. La racine est obtenue par élimina-

radical (*i.e.* une des formes prises par la racine). Il existe, là aussi, différentes approches pour prendre en compte ce type de morphologie, comme le recours à des ressources morphologiques spécifiques ou l'utilisation d'analyseurs dérivationnels, *i.e.* des analyseurs morphologiques s'appliquant aux formes dérivées (Daille *et al.*, 2002). Enfin, d'autres approches moins sophistiquées, considérées généralement comme des approximations des traitements linguistiques, permettent de traiter la variation morphologique des unités lexicales. Les différentes variantes d'un même mot sont alors regroupées autour d'un *stem* (ou pseudo-racine) qui se rapproche de la notion de racine (sans nécessairement avoir une origine étymologique comme dans le cas de la racine linguistique). Le traitement associé est la procédure de *stemming* (ou racinisation) qui présente l'avantage d'être moins complexe à mettre en œuvre que les deux types de traitements précédents. Elle prend en compte à la fois les cas relevant des morphologies flexionnelle et dérivationnelle.

Ces quelques notions fondamentales de morphologie ayant été définies, nous nous intéressons plus précisément au traitement de la variation morphologique en RI.

2.2.2 Traitement de la variation morphologique en RI

Comme mentionné en introduction de ce chapitre, il existe principalement deux manières de prendre en compte les variantes morphologiques au sein d'un SRI. La première consiste à appliquer un traitement morphologique par *stemming* ou à l'aide d'un analyseur flexionnel ou dérivationnel lors de la phase d'indexation des documents et requêtes. Il s'agit d'une approche par conflation. Les différentes variantes possibles d'un mot sont normalisées, *i.e.* ramenées à une seule et même forme, (pseudo-)racine ou lemme, lors de l'indexation. L'appariement des documents et de la requête se fait alors sur la base de cette forme canonique. La seconde approche est de recourir aux traitements morphologiques pour la tâche d'extension des requêtes (approche par extension). Il s'agit de procéder également à la reconnaissance des variantes morphologiques sans opérer cependant de traitement de normalisation. Les termes de la requête de l'utilisateur sont enrichis par le biais de leurs variantes morphologiques au moment de la recherche. Les expériences présentées dans ce chapitre visent principalement à prendre en compte les variantes morphologiques lors de l'indexation des documents et requêtes. Les approches par extension seront abordées plus en détail au chapitre 5 de ce manuscrit.

Les différentes expérimentations réalisées pour évaluer l'intérêt de recourir à un niveau d'analyse morphologique des documents et requêtes en RI n'utilisent pas nécessairement les mêmes types de traitements morphologiques. Afin d'en dresser un bilan efficace, nous décomposons leur description selon le type d'outils qu'elles manipulent. Nous présentons tout d'abord les expériences qui exploitent une procédure de *stemming* en RI, puis celles qui utilisent des analyseurs morphologiques flexionnels et/ou dérivationnels.

---

tion de tous les affixes et désinences. Elle est porteuse des sèmes essentiels, communs à tous les termes constitués avec cette racine » (Larousse, 1998).

2.2.2.1 Impact du *stemming*

Les outils utilisés pour procéder à la racinisation des mots des documents et requêtes, les *stemmers*, reposent généralement sur une liste d'affixes de la langue considérée et sur un ensemble de règles de désuffixation construites *a priori*, qui permettent, étant donné un mot, de retrouver son *stem*. Les *stemmers* traditionnellement utilisés en RI pour l'anglais sont ceux de Porter (1980) et Lovins (1968).

De nombreux travaux se sont intéressés à l'utilisation du *stemming* en RI, mais son impact sur les performances de SRI est cependant très variable selon les expérimentations. Les expériences de Lennon *et al.* (1981) et Harman (1991) mesurant l'influence de la racinisation pour l'anglais aboutissent à des conclusions globalement décevantes puisqu'aucune amélioration de résultats n'est constatée par rapport à un SRI « traditionnel ». Les mêmes observations sont obtenues par Fuller et Zobel (1998), qui comparent l'apport respectif de quatre types de *stemmers* (les algorithmes de Porter, de Lovins, un *stemmer* qui supprime simplement les marques du pluriel et un *stemmer* à base de dictionnaires). Leur conclusion est que les améliorations apportées par ce traitement restent insuffisantes, bien que réelles pour certaines requêtes. Hull (1996), lors de nouvelles expérimentations, tente de justifier ces faibles résultats en montrant que les mesures d'évaluation traditionnelles de la RI (*i.e.* le rappel et la précision) ne sont pas forcément appropriées pour évaluer précisément l'influence du *stemming*. En intégrant notamment la prise en compte de la longueur des requêtes et du nombre de documents pertinents retournés pour chacune d'elles, et en s'appuyant sur des collections de documents plus grandes, il démontre que le *stemming* est efficace pour l'anglais (une hausse comprise entre 1 et 3% par rapport à des systèmes ne prenant pas en compte les variantes morphologiques), sauf pour les requêtes longues. Cette amélioration n'est toutefois pas effective sur de très petites collections de documents. Les expériences de Krovetz (1993) apparaissent également encourageantes puisque la racinisation conduit à une hausse des résultats située entre 1,3% et 45,3% selon les collections et les techniques de *stemming* utilisées. Les améliorations les plus conséquentes sont obtenues dans le cas de documents courts (environ 45 mots) associés à des requêtes courtes (comportant 7 mots en moyenne).

Il ressort donc de ces différentes expériences que l'influence du *stemming* en RI est tributaire d'un certain nombre de facteurs. Outre la longueur des requêtes et la taille des collections déjà mentionnées, les travaux étudiés en distinguent principalement deux. Le premier est étroitement lié à la qualité du *stemmer* utilisé. Une mauvaise racinisation, provoquée par des erreurs de sur-racinisation (*e.g.* la pseudo-racine *nat* qui regroupe à la fois *nature* et *nation*) ou de sous-racinisation (*e.g.* la pseudo-racine *adaptat* qui empêche le regroupement des formes *adapter* et *adaptation*), a en effet pour conséquence de regrouper des variantes qui font référence à des concepts différents, ce qui entraîne une dégradation des performances d'un SRI l'utilisant. Pour contrôler le processus de racinisation et réduire le nombre d'erreurs engendrées, plusieurs stratégies ont été proposées. La première consiste à coupler à un *stemmer* traditionnel des dictionnaires électroniques. Leur rôle est de stopper le processus de suppression de suffixes lorsque le mot obtenu y est trouvé. Bien que l'utilisation de *stemmers* à base de diction-

naires règle en grande partie les problèmes des algorithmes de *stemming* traditionnels, les performances obtenues à la suite de différentes expérimentations (Krovetz, 1993; Fuller et Zobel, 1998) démontrent que ces outils fournissent généralement une bonne précision mais passent à côté de beaucoup de termes (*i.e.* des termes qui auraient dû être rapprochés car étant morphologiquement liés) absents de la ressource. La seconde stratégie adoptée pour vérifier que les variantes regroupées possèdent effectivement un lien morphologique consiste à s'appuyer sur les cooccurrences. La méthode proposée par Xu et Croft (2000) par exemple, qui fonctionne en corpus, sans ressources ni règles prédéfinies, consiste dans un premier temps à normaliser les mots à l'aide d'un *stemmer* de type Porter, puis à rassembler au sein d'une même classe d'équivalence morphologique (*e.g.* *stocks*, *stock*, *stocked*, *stocking*...) les mots racinisés qui, en plus de posséder la même pseudo-racine, cooccurrent de façon significative (valeur mesurée à l'aide d'une variante de l'information mutuelle). L'utilisation des cooccurrences présuppose l'idée d'une proximité sémantique entre termes. Les résultats obtenus en utilisant ces classes de mots pour l'extension de requêtes montrent une amélioration significative des performances. D'une manière générale, il semble, à travers ces différentes observations, que la procédure de *stemming* doive être couplée à des traitements complémentaires si l'on souhaite qu'elle ait un impact plus important en RI. Une façon d'y parvenir est de ne pas se contenter d'utiliser des ressources construites *a priori*, comme c'est le cas pour la plupart des *stemmers* utilisés dans les expériences citées plus haut, mais de s'appuyer, au contraire, sur des connaissances morphologiques acquises automatiquement à partir des documents et requêtes analysés. Nous reviendrons plus en détail sur cette idée dans le chapitre 5. Le second facteur qui entre en jeu dans l'évaluation précise que l'on peut faire de l'impact du *stemming* est la langue sur laquelle il est appliqué. Comme le soulignent Arampatzis *et al.* (2000), l'efficacité du *stemming* dépend de la complexité morphologique de la langue. Les expériences réalisées sur des langues morphologiquement plus riches que l'anglais, comme par exemple le slovène (Popovic et Willett, 1992) ou le suédois (Carlberger *et al.*, 2001) montrent que la procédure de *stemming* est intéressante en RI, puisqu'une augmentation du rappel et de la précision, comprise entre 15 et 40% selon les expérimentations et les langues considérées, est observée pour des textes d'une longueur moyenne. Il apparaît donc que plus une langue est morphologiquement riche, plus il devient pertinent de prendre en compte le niveau morphologique à travers une procédure de *stemming*, voire par le biais de traitements morphologiques plus évolués.

Afin d'avoir une vision générale de l'intérêt des informations morphologiques au sein des SRI, nous nous intéressons maintenant aux systèmes qui intègrent une analyse morphologique plus sophistiquée des documents et requêtes, analyse qui s'effectue par le biais de lemmatiseurs et d'analyseurs dérivationnels. Notre choix de ne pas faire de distinction dans notre présentation entre ces deux types d'outils se justifie par le fait que, au sein des diverses expériences que nous allons décrire, ces traitements sont souvent couplés en raison de leur complémentarité.

2.2.2.2 Impact d'analyseurs morphologiques flexionnels et dérivationnels

Nous présentons ici quelques résultats d'expérimentations réalisées pour évaluer l'impact des informations morphologiques obtenues par le biais de ces analyseurs sur les performances des SRI, puis proposons quelques points-clés liés à l'application de ces traitements en RI.

Quelques résultats

Les expériences de Gaussier *et al.* (1997) pour le français montrent l'apport de la morphologie flexionnelle en RI. Les résultats obtenus lors de l'intégration d'un module de lemmatisation[4] dans leur SRI présentent une amélioration de la précision moyenne de 16%. Ces auteurs proposent également de combiner le traitement de lemmatisation à un module dit de morphologie relationnelle[5]. Par le biais d'une méthode d'apprentissage non supervisée des suffixes et des opérations de suffixation à partir de lexiques flexionnels de la langue, le système proposé permet l'extraction de suffixes potentiels ensuite utilisés pour le regroupement de variantes. Ce traitement additionnel offre une augmentation complémentaire de la précision moyenne de 2%.

Zweigenbaum *et al.* (2001) montrent également l'apport faible mais réel de la lemmatisation et de la dérivation dans une tâche d'appariement entre requêtes et termes normalisés, appliqués plus précisément à un domaine de connaissance restreint (le domaine médical). L'utilisation de connaissances flexionnelles[6] et dérivationnelles améliore en moyenne les réponses à une requête. La flexion agit dans 6,6% des cas avec une hausse modeste, et la dérivation agit dans 2% des cas avec une augmentation plus nette.

Les expériences de Vilares-Ferro *et al.* (2002) pour l'espagnol exploitent également successivement ces deux types d'analyseurs. La première étape du système consiste à étiqueter morpho-syntaxiquement les unités lexicales et à obtenir les lemmes des textes à indexer. Chacun des lemmes est ensuite remplacé par le représentant de la famille morphologique à laquelle il appartient. Une hausse significative du rappel peut alors être constatée par rapport à une procédure de *stemming* traditionnelle.

Points-clés

Au vu de ces expériences, il apparaît que l'analyse morphologique flexionnelle et dérivationnelle des documents et requêtes peut être considérée comme une tâche pertinente en RI. Grâce aux règles linguistiques fortes qu'elle met en jeu, elle limite les rapprochements de termes non liés. De plus, son application peut constituer une première étape de désambiguïsation des mots, l'analyse et la production de la forme de base d'un terme polysémique pouvant en effet nécessiter la détermination automatique de sa catégorie morpho-syntaxique (*e.g.* *porte* = nom ou verbe) en se

[4] Utilisation des outils de lemmatisation de Xerox.

[5] Nous considérons ici le terme relationnel comme équivalent à dérivationnel. Pour une distinction plus précise de ces notions, se référer à (Gaussier *et al.*, 1997).

[6] L'analyseur flexionnel utilisé est le lemmatiseur FLEMM développé par Namer (2000).

basant sur son contexte d'apparition. La prise en compte des variantes allomorphiques (*i.e.* un morphème peut avoir plusieurs formes ; *e.g.* *(tu) paies ou payes*) favorise le rappel. Néanmoins, comme pour le *stemming*, l'impact d'une analyse morphologique (flexionnelle ou dérivationnelle) semble étroitement lié à la langue prise en compte. Les expériences présentées ici ont toutes été menées sur des langues morphologiquement riches.

La combinaison des deux types de morphologie paraît être une solution efficace en vue de l'accroissement des performances des systèmes. L'intérêt du couplage est justifié, en effet, par la complémentarité des deux analyses : la lemmatisation permet, sans générer un nombre d'erreurs important, le regroupement de variantes morphologiques de même catégorie grammaticale. L'analyse dérivationnelle conduit ensuite, à partir des résultats de la lemmatisation, à rassembler (de façon plus précise qu'une procédure de *stemming* car motivée linguistiquement) les variantes morphologiques quelle que soit leur catégorie grammaticale. Savoy (1993) a toutefois souligné l'importance de bien cloisonner les traitements liés à chacune de ces morphologies afin d'optimiser au mieux leur efficacité.

2.2.3 Bilan de l'apport de connaissances morphologiques en RI

Dans l'ensemble, ces différentes expériences tendent à montrer l'intérêt de prendre en compte les variantes morphologiques pour améliorer le rappel et la précision des SRI. L'apport de ces connaissances semble toutefois fortement lié, nous l'avons vu, à plusieurs facteurs comme le type de collection utilisé (longueur des requêtes, taille des documents...) ou la langue prise en compte. Il dépend également de la qualité des informations exploitées et, par conséquent, de celle des outils utilisés pour leur mise en évidence à partir des textes, qui doivent notamment s'assurer que les différentes variantes regroupées possèdent véritablement un lien morphologique. Concernant ces outils (*stemmer* ou analyseurs flexionnels et dérivationnels), il reste toutefois difficile de choisir clairement celui qui semble réellement le plus adapté. Plusieurs expériences ont cherché à comparer leur efficacité. Celles de (Hull et Grefenstette, 1996) montrent une légère supériorité de l'analyse flexionnelle pour l'anglais, mettant toutefois en évidence une variabilité importante des résultats selon les requêtes prises en compte. Les expérimentations de (Moulinier *et al.*, 2000) qui comparent ces deux types d'outils sur des collections en anglais et en français donnent également l'avantage à la lemmatisation mais uniquement pour le français. D'une manière plus générale, les résultats observés rejoignent les conclusions évoquées précédemment pour chacun des outils. La procédure de *stemming* semble bénéfique à certaines conditions (*e.g.* selon la langue considérée ou le type de *stemmer* appliqué) ; l'approche la plus efficace, nous l'avons vu, consiste à coupler un racineur avec une procédure de contrôle pour limiter le nombre d'erreurs engendrées. L'analyse morphologique d'ordre flexionnel, qui paraît d'ailleurs assez fiable, contribue également à améliorer les performances des systèmes, et semble plus particulièrement adaptée aux langues morphologiquement riches. Une de ses limites réside toutefois dans l'utilisation de ressources construites *a priori*, qui ne lui permettent pas de prétendre à une couverture exhaustive. Son couplage avec un traitement de morpho-

logie dérivationnelle, en procédant au rassemblement des formes indépendamment de leurs catégories grammaticales, offre l'avantage de retrouver plus de variantes.

Plus généralement, l'apport réel des informations morphologiques sur les performances des SRI pourrait certainement être accru par l'intégration d'analyses appartenant aux autres niveaux de langue (*e.g.* exploitation du niveau syntaxique ou sémantique). Un certain nombre d'expériences montrent que des erreurs des traitements morphologiques sont liées à la non prise en compte des termes complexes (extractibles par une analyse syntaxique) ou la présence de termes morphologiquement ambigus (désambiguïsables par le biais d'informations sémantiques ou syntaxico-sémantiques).

2.3 Apport de connaissances syntaxiques en RI

Le principal intérêt de recourir à des informations syntaxiques en RI est d'une part de dépasser la notion de chaînes de caractères traditionnellement utilisée en RI en prenant en compte notamment des unités lexicales composées de plusieurs termes (*e.g.* *effet de serre*) plus significatives et moins ambiguës et, d'autre part, de pallier les limites des représentations en « sacs de mots » en considérant les relations et dépendances susceptibles d'exister entre les termes. Dans le but d'avoir une idée précise de l'apport de ces connaissances en RI, nous nous intéressons ici aux résultats d'expériences qui les intègrent au sein de systèmes. Après un bref rappel de quelques notions utiles de syntaxe (section 2.3.1), nous présentons (section 2.3.2) les diverses expérimentations qui exploitent ces informations en RI, et dressons un bilan de leur impact. Nous terminons en évoquant (section 2.3.3) les différentes adaptations nécessaires aux SRI souhaitant manipuler de telles connaissances.

2.3.1 Quelques notions utiles de syntaxe

La syntaxe s'intéresse à la structure des phrases et des syntagmes (*i.e.* séquences de mots formant des unités syntaxiques (Riegel *et al.*, 1999)). L'analyse syntaxique des documents permet notamment, selon l'application visée, de prendre en compte l'ordre des mots dans une phrase, d'identifier les fonctions grammaticales des termes (*e.g.* reconnaître le sujet ou le complément d'objet direct d'une phrase...), de lever les ambiguïtés des termes (*e.g.* déterminer si le mot *avions* est dans le contexte de la phrase un nom ou un verbe), d'identifier les termes complexes (*e.g.* *petits pois*)...

On distingue traditionnellement deux types d'analyse syntaxique. L'analyse des phrases en constituants d'une part cherche, en s'appuyant par exemple sur des règles affirmant qu'une phrase est composée d'un syntagme nominal et d'un syntagme verbal, qu'un syntagme nominal peut lui même être formé d'un nom propre ou d'un déterminant suivi d'un adjectif et d'un nom commun... à produire un arbre syntaxique d'analyse complète de la phrase ou au moins des parties traitables. L'analyse en dépendances, d'autre part, vise à mettre en évidence les relations de dépendance entre les mots « têtes » (éléments centraux de syntagmes) et les mots « modifieurs » qu'ils régissent, que ce soit globalement au niveau de la phrase (où le prédicat verbal constitue fréquemment le mot tête principal régissant en particulier le sujet) ou à l'intérieur d'un syntagme.

Ainsi, selon Strzalkowski *et al.* (1999), les groupes de mots *information retrieval, retrieval of information, retrieve more information* et *information that is retrieved*... peuvent être ramenés à la relation *retrieve+information* où *retrieve* est l'élément tête et *information* son modifieur.

2.3.2 Exploitation d'informations syntaxiques en RI

Bien que l'analyse syntaxique permette, comme nous l'avons vu, d'extraire des textes différents types d'informations syntaxiques, l'exploitation de ces connaissances en RI se limite généralement à la prise en compte de syntagmes. Après avoir présenté (en section 2.3.2.1) comment ceux-ci pouvaient être utilisés en RI, nous nous intéressons (en section 2.3.2.2) à leur impact sur les performances des systèmes à travers une synthèse des résultats d'expérimentations réalisées en ce sens.

2.3.2.1 Exploitation de syntagmes en RI

Les syntagmes exploités en RI peuvent être de différentes formes selon le type d'analyse syntaxique appliquée aux textes. Nous présentons tout d'abord les deux principaux types de syntagmes généralement pris en compte puis nous nous intéressons à leur intégration au sein des systèmes.

Les syntagmes en RI

Les syntagmes peuvent être des termes complexes, que nous considérons ici comme toute unité lexicale constituée d'au moins deux mots pleins[7], auxquels peuvent s'adjoindre des déterminants et des prépositions (en français par exemple la structure *Nom Prép (Det) Nom* (*e.g.* *pommes de terre*)). Les termes complexes se substituent alors aux termes simples en tant que termes d'indexation. Leur principal avantage par rapport aux termes simples est qu'ils sont moins ambigus, et souvent plus aptes à désigner des concepts puisqu'ils réfèrent généralement un domaine de connaissance spécialisé. De ce fait, leur intégration au sein des SRI doit permettre d'améliorer leur précision. L'extraction de ces termes ne nécessite pas obligatoirement une analyse syntaxique très poussée des textes. Certains techniques automatiques d'acquisition vont même jusqu'à exploiter uniquement des indices numériques (aspect fréquentiel). Ces approches numériques, simples à mettre en œuvre mais parfois imprécises, repèrent des séquences de mots qui apparaissent ensemble plus fréquemment que le hasard ne l'aurait permis. D'autres techniques plus poussées proposent, en s'appuyant sur des indices structurels (aspect symbolique), voire sur une combinaison d'approches numériques et symboliques (Claveau, 2003; Claveau et Sébillot, 2004a), d'extraire des combinaisons de mots à structure syntagmatique connue (*e.g.* la structure *Nom Prép*

[7] D'un point de vue terminologique, la définition des termes complexes est souvent plus compliquée et fait l'objet de nombreuses discussions. Par exemple, une distinction est souvent faite entre les syntagmes lexicalisés qui figent une construction syntaxique (*e.g.* *fil de fer barbelé*) et les mots formés par composition (*e.g.* *bébé-éprouvette*) (Riegel *et al.*, 1999). Dans certaines approches, seuls les premiers correspondent véritablement aux termes complexes.

(Det) Nom citée précédemment), qui donnent souvent des résultats plus pertinents pour la RI (Hull *et al.*, 1997).

Outre l'extraction de termes complexes, certains travaux vont plus loin et proposent de prendre en compte également leurs variantes. En effet, ces termes, plus encore que les termes simples, sont sujets à de multiples variations. On distingue généralement (Daille, 1996, 2002; Jacquemin, 1994) les variantes typographiques (*e.g.* *système expert* et *système-expert*), morpho-syntaxiques (*e.g.* *ulcère de la cornée* et *ulcère cornéen*), syntaxiques (*e.g.* *sécurité des réseaux* et *sécurité des données et des réseaux*)... Ces variantes, détectées par le biais d'une analyse syntaxique plus fine, sont alors normalisées, *i.e.* ramenées à une forme unique qui est alors utilisée comme terme d'indexation. Les travaux de Jacquemin *et al.* (1997) démontrent la nécessité de prendre en compte les variations des termes complexes pour une utilisation en RI.

En procédant à une analyse syntaxique plus sophistiquée (*e.g.* une analyse syntaxique en dépendances), il est possible de mettre en évidence les relations de dépendances entre les mots présents au sein d'un syntagme, et de dégager plus précisément des relations tête+modifieur (*cf.* l'exemple de *information retrieval* cité précédemment). La normalisation des syntagmes en structure tête+modifieur présente l'avantage de regrouper les différentes variantes en une seule et même forme. L'impact des relations tête+modifieur sur les performances des SRI qui les exploitent est fortement tributaire, nous le verrons à travers les expériences de l'état de l'art, de la qualité de l'analyse syntaxique mise en œuvre.

Bien que les notions de termes complexes et de syntagmes structurés en tête+modifieur soient liées[8], nous les distinguons lors de la description des expériences ci-dessous (section 2.3.2.2).

Intégration au sein de SRI

D'une manière générale, les syntagmes (*i.e.* les termes complexes et les syntagmes structurés en tête+modifieur), comme les connaissances morphologiques évoquées précédemment, peuvent être pris en compte au sein des SRI de deux manières. Lors de la phase d'indexation, les termes complexes (ramenés à une forme unique si leurs variantes sont prises en compte) ou les syntagmes normalisés en tête+modifieur sont utilisés comme termes d'indexation. L'appariement entre les documents et requêtes se fait alors sur cette base et devrait permettre de retrouver davantage de documents qu'une mise en correspondance à l'aide de termes simples puisque les documents contenant uniquement les variantes (des termes complexes ou des syntagmes structurés) doivent également être retournés. Pour les structures tête+modifieur, l'objectif est également de favoriser les documents dans lesquels les termes simples formant le syntagme entretiennent la même relation de dépendance que dans la requête. Ainsi, si dans la question de l'utilisateur le terme *logiciel* est en position modifieur au sein d'un syntagme (*e.g.* *installation de logiciel*), il est alors possible de

[8]Les termes complexes peuvent être considérés comme un sous-ensemble particulier de syntagmes structurés en relation tête+modifieur.

laisser de côté tous les documents où *logiciel* est en position tête (*e.g.* dans *logiciel de comptabilité* ou *logiciel client*). En expansion de requêtes, il convient d'identifier toutes les variantes des syntagmes qu'elles contiennent et de les enrichir par elles.

Nous présentons à présent les résultats de différentes expériences intégrant ces syntagmes en RI.

2.3.2.2 Résultats de l'exploitation de syntagmes en RI

Nous nous intéressons ici tout d'abord aux expériences qui intègrent au sein de SRI des termes complexes puis à celles qui exploitent des syntagmes structurés en tête+modifieur.

Exploitation de termes complexes

Un certain nombre de travaux ont cherché à prendre en compte des termes complexes au sein de SRI. Les résultats obtenus sont globalement mitigés et, comme pour les informations morphologiques, l'impact de ces informations sur les performances des SRI est lié à un certain nombre de facteurs et dépend notamment des méthodes d'extraction (numérique ou symbolique) des termes complexes retenus.

Pour ce qui concerne les termes complexes « statistiques » (*i.e.* obtenus par le biais d'une approche numérique), plusieurs expériences (Salton, 1975; Fagan, 1987) montrent que l'intégration de termes complexes extraits par repérage de cooccurrences lors de la phase d'indexation améliore la précision moyenne des systèmes par rapport à des SRI n'utilisant que des termes simples. Ces améliorations sont toutefois tributaires de plusieurs paramètres. Nous en distinguons ici principalement trois. Le premier concerne, comme le montre Fagan (1987), la taille des collections utilisées lors des expérimentations. Plus le nombre de documents est élevé, plus les méthodes numériques d'extraction sont fiables puisqu'elles reposent essentiellement sur la notion de fréquence et plus les termes extraits ont de chances d'être sémantiquement pertinents. Le deuxième est lié au domaine de la base documentaire utilisée. Les termes complexes, fortement représentés en domaine spécialisé, sont généralement moins fréquents dans les collections généralistes et tendent à ne pas être extraits par des méthodes d'acquisition numériques. Le dernier est lié à la langue de la collection. Les expériences de Gaussier *et al.* (2000) sur le français par exemple — réalisées sur un corpus d'évaluation de petite taille — ne permettent pas de prouver l'intérêt des termes complexes en RI.

L'apport de termes complexes « syntaxiques » (*i.e.* acquis à l'aide de méthodes symboliques), théoriquement plus fiables, est également très variable. Les résultats obtenus à la suite des expériences relatées dans (Fagan, 1987; Lewis, 1992) ou (Gaussier *et al.*, 2000) pour le français ne montrent aucune amélioration significative par rapport à une indexation par termes simples, et sont même en-dessous de ceux obtenus grâce à des termes « statistiques ». Une des raisons pouvant justifier ces faibles résultats est liée aux conditions d'évaluation utilisées qui s'avèrent peu optimales (nombre de requêtes considéré trop faible et tailles des collections trop petites). D'autres expériences (Dillon et Gray, 1983; Hull *et al.*, 1997) sont plus positives et prouvent que, dans certains cas,

l'intégration de termes complexes « syntaxiques » dans un SRI surpasse une indexation à base de termes simples ou de termes complexes « statistiques ».

D'une manière plus générale, à travers ces diverses expériences, il est difficile de conclure de manière tranchée sur l'intérêt ou non de recourir à des termes complexes en RI, voire sur la méthode d'acquisition privilégiée de ces termes. Il est évident que leur efficacité est étroitement liée à la qualité des éléments recueillis, mais dépend également de la représentation de ces termes dans les index. Nous le verrons, les mesures de pondération utilisées pour mettre en valeur leur pouvoir de représentativité des contenus textuels doivent nécessairement être adaptées au fait que ces termes sont généralement plus faiblement fréquents dans les documents.

Exploitation de syntagmes structurés en relation tête+modifieur

L'exploitation de syntagmes structurés en tête+modifieur, généralement pris en compte en complément des termes simples, a fait l'objet de plusieurs expériences. Les travaux de (Strzalkowski *et al.*, 1999; Haddad, 2003; Arampatzis *et al.*, 1996) attestent tout d'abord de l'intérêt de prendre en compte ces structures pour l'anglais puisqu'une amélioration significative des performances des SRI aussi bien en termes de précision que de rappel est constatée par rapport à une indexation par termes simples (augmentation comprise entre 5 et 30% selon les expérimentations). Les hausses les plus importantes sont obtenues sur des requêtes longues, susceptibles de contenir davantage de syntagmes en relation tête+modifieur. Ces travaux diffèrent par le type d'analyse syntaxique mis en œuvre (les outils syntaxiques utilisés ne s'appuient pas nécessairement sur les mêmes types de grammaire), les méthodes utilisées pour filtrer les syntagmes extraits des documents et ne retenir que les plus pertinents (Haddad (2003) opère un filtre syntaxique, Arampatzis *et al.* (1996) s'appuient sur des traitements numériques et Strzalkowski *et al.* (1999) ne procèdent à aucune sélection)[9], et les types de syntagmes retenus (Arampatzis *et al.* (1996); Strzalkowski *et al.* (1999) par exemple prennent en compte des syntagmes nominaux et verbaux). Des expériences similaires ont été menées sur d'autres langues, comme celles de Kraaij et Pohlmann (1996) pour l'allemand, de Vilares-Ferro *et al.* (2002) pour l'espagnol ou de Fajah *et al.* (1996) pour le français, et révèlent une amélioration systématique des performances. Enfin, d'autres expériences ont cherché à évaluer l'utilisation des relations tête+modifieur non plus au niveau de l'indexation mais en post-traitement des SRI, dans le but d'améliorer les performances obtenues par un premier SRI basé sur des méthodes traditionnelles d'indexation par termes simples (Mitra *et al.*, 1997). Les résultats observés ne montrent cependant pas d'amélioration significative.

À la suite de ces expériences, il semble globalement intéressant de prendre en compte les syntagmes structurés en tête+modifieur au sein des SRI puisqu'une hausse systématique des performances des systèmes est observée (*cf.* également les travaux de Zhai *et al.* (1997) confirmant ces conclusions). L'exploitation de ces structures est toutefois

[9]Certains syntagmes, comme par exemple *l'évolution du processus d'apprentissage de lecture*, sont difficilement décomposables ; un filtrage (selon une approche numérique ou symbolique) s'avère dans ce cas utile.

fortement sensible à la longueur des requêtes : une requête courte pourra avoir peu de chances de contenir des syntagmes de ce type, rendant par conséquent leur prise en compte inutile.

Globalement, l'exploitation de syntagmes en RI peut s'avérer intéressante, sous couvert, comme nous venons de le voir, que ceux-ci soient extraits et utilisés dans de bonnes conditions. Nous avons limité cette synthèse aux termes complexes et aux syntagmes de type tête+modifieur. D'autres types de structures syntaxiques ont été prises en compte en RI (Smeaton, 1999; Metzler et Hass, 1989; Matsumura *et al.*, 2000), obtenues par le biais d'une analyse syntaxique complète de la phrase visant à rendre compte de dépendances plus complexes entre les termes que les relations tête+modifieur (*e.g.* les dépendances entre syntagmes). L'intérêt de l'exploitation de ces structures en RI n'a toutefois pas été démontré.

L'apport des syntagmes est cependant étroitement lié à la façon dont les SRI sont adaptés à leur prise en compte. Nous revenons donc à présent sur les modifications qu'il est nécessaire d'effectuer au cœur même des systèmes pour pouvoir exploiter pleinement la richesse de ces connaissances syntaxiques.

2.3.3 Adaptation des SRI pour l'intégration d'informations syntaxiques

Comme nous l'avons évoqué dans le chapitre précédent, les SRI, initialement conçus pour prendre en compte des termes simples, ne sont pas toujours adaptés pour accueillir des informations linguistiques poussées. Parmi les différents mécanismes traditionnels de RI, deux doivent plus particulièrement être modifiés pour la prise en compte des syntagmes. Le premier concerne les mesures de pondération des termes utilisées pour déterminer leur pouvoir de représentativité du contenu textuel. Ces mesures s'appuient, nous l'avons vu, essentiellement sur la notion de fréquence. Or, ces syntagmes, bien que fortement porteurs de sens, sont généralement beaucoup moins fréquents que les termes simples et sont alors sous-pondérés. Plusieurs travaux ont cherché à établir de nouvelles mesures de pondération. Certains proposent de pondérer l'expression (terme complexe) en fonction du poids de ses composantes, les résultats obtenus n'étant toutefois pas uniformes (Fagan, 1987; Lewis et Croft, 1990). D'autres (Haddad, 2002; Pedersen et Bruce, 1997; Arampatzis *et al.*, 1996) ont recours à une pondération dite « syntaxique », généralement basée sur les catégories grammaticales des constituants du syntagme, en accordant plus d'importance par exemple à certains types de syntagmes (comme les nominaux) (Pedersen et Bruce, 1997) ou en favorisant leur élément « tête » (Arampatzis *et al.*, 1996). Les résultats obtenus montrent globalement une amélioration des performances des SRI. L'impact des termes complexes ou des structures tête+modifieur en RI est donc étroitement lié aux mesures utilisées pour leur pondération. Comme le remarque Spärck Jones (1999), une mauvaise pondération de ces structures complexes peut rapidement devenir néfaste au processus de recherche.

Le deuxième facteur à prendre en compte est la façon dont ces syntagmes sont intégrés au sein des index, et leur combinaison avec les termes simples. Dans le modèle vectoriel, deux stratégies d'intégration ont été principalement expérimentées. La première

consiste à séparer l'indexation des syntagmes de celle des termes simples. La technique utilisée (Fox, 1983) substitue à une représentation vectorielle classique, regroupant l'ensemble des termes d'indexation au sein d'un seul vecteur, une représentation en deux sous-vecteurs différents. Cette représentation peut également être utilisée pour l'expansion de requêtes, ces dernières étant enrichies avec des termes provenant de ces deux parties. Les différentes expérimentations (Haddad, 2002; Kraaij et Pohlmann, 1996) montrent une amélioration de la précision des systèmes. La seconde approche (Strzalkowski *et al.*, 1999; Arampatzis *et al.*, 2000) privilégie l'utilisation d'un index unique pour les termes simples et les syntagmes. Ces derniers sont utilisés en complément des termes simples. L'intérêt de cette méthode par rapport à la précédente est de pouvoir s'appuyer sur les composantes (*i.e.* les termes simples) des syntagmes pour mettre en correspondance un document et une requête, au cas où ces syntagmes ne peuvent être utilisés pour l'appariement.

2.3.4 Bilan de l'apport de syntagmes en RI

En conclusion, la prise en compte des syntagmes, en offrant une description plus riche du contenu informationnel, semble pertinente dans un cadre de RI. Leur impact dépend cependant tout d'abord de la qualité des informations extraites des documents et requêtes. Pour les termes complexes en particulier, cette conclusion tend à privilégier les méthodes symboliques, susceptibles d'acquérir des informations syntaxiques plus fiables. Les techniques du TAL, à condition d'être suffisamment souples, présentent donc une valeur ajoutée par rapport aux mécanismes traditionnels de RI. L'apport des syntagmes est aussi fortement lié à la manière dont ils sont intégrés au sein des SRI, et aux adaptations opérées au cœur même des modèles pour leur prise en compte. Enfin, l'évaluation précise de leur efficacité sur les résultats des systèmes dépend des conditions d'expérimentations utilisées (types de collections (nombre de documents, domaine), de requêtes (longueur, nombre), langue prise en compte...).

2.4 Apport de connaissances sémantiques en RI

Le recours à une analyse sémantique des documents et requêtes en RI vise à extraire des informations sur le sens des mots et sur les relations que ces mots entretiennent entre eux. Ces informations ont pour objectif d'une part, en identifiant par exemple que les termes *vélo* et *bicyclette* sont liés par une relation de synonymie ou que les termes *écrevisse* et *crustacé* sont en relation d'hyperonymie[10], d'offrir la possibilité aux SRI de mettre en correspondance des termes sémantiquement proches bien que graphiquement différents. Elles visent d'autre part à rendre plus pertinent le processus d'appariement en permettant par exemple d'identifier que le sens du terme *avocat* utilisé dans la requête d'un utilisateur (qui fait référence par exemple à la profession juridique) est différent de celui de la même chaîne graphique contenue dans le document (qui renvoie au fruit). Nous revenons dans un premier temps sur les divers types de connaissances séman-

[10] L'hyperonyme est un incluant, un synonyme à un niveau de généralité supérieur.

tiques utilisées dans un cadre de RI (section 2.4.1). Nous nous intéressons ensuite à leur intégration au sein des SRI et à l'évaluation de leur apport (section 2.4.2), en distinguant leur utilisation en extension de requêtes et pour l'indexation des documents et requêtes. Un problème récurrent à toutes ces études étant lié au caractère polysémique des termes, nous terminons (section 2.4.3) par une présentation des principaux travaux de désambiguïsation automatique utilisés en RI.

2.4.1 Informations sémantiques exploitées en RI

Les connaissances sémantiques qui sont intégrées au sein des SRI peuvent avoir plusieurs origines. Elles peuvent tout d'abord être issues de bases lexicales existantes, dont la plus connue est WORDNET (Fellbaum, 1998). Cette ressource présente la particularité de couvrir la majorité des mots (noms, verbes, adjectifs et adverbes) de la langue anglaise et de rendre compte des relations sémantiques (*e.g.* synonymie, hyperonymie, méronymie[11], antonymie...) qu'ils entretiennent. L'exploitation des informations issues de cette ressource peut s'avérer, nous le verrons, intéressante en RI. Une telle base généraliste pose néanmoins le problème de l'adéquation des informations contenues avec le ou les domaines précis de la collection de documents. Pour s'affranchir de cette limite, une solution possible est d'utiliser des informations sémantiques acquises automatiquement à partir des collections de textes (Claveau, 2003). L'extraction de ces connaissances en corpus peut être effectuée par le biais de méthodes numériques (*i.e.* basées sur des indices numériques). Ces approches permettent par exemple de détecter des associations d'unités lexicales (ou cooccurrences ; *cf.* section 2.3.2.2), pouvant correspondre à des termes complexes ou à des unités en relations syntagmatiques (Grefenstette, 1994). Elles offrent, d'autre part, la possibilité, en procédant à une analyse des mots qui partagent les mêmes propriétés contextuelles (*e.g.* les mêmes contextes syntaxiques, ou les mêmes mots cooccurrant dans une certaine fenêtre (Pichon et Sébillot, 2000)), de faire émerger des classes à caractère conceptuel d'unités lexicales et de découvrir des relations paradigmatiques (*e.g.* synonymie, hyperonymie...) entre ces unités. Les connaissances sémantiques peuvent également être acquises en corpus à partir d'indices structurels — on parle d'une approche symbolique de l'acquisition. Divers travaux (Jouis, 1995; Hearst, 1992; Morin, 1999; Claveau, 2003) ont ainsi extrait, en s'appuyant sur une expertise linguistique ou des techniques d'apprentissage, des relations sémantiques (telle que l'hyperonymie par exemple).

Plusieurs expériences ont cherché à exploiter de telles connaissances en RI. Elles font l'objet de la section suivante.

2.4.2 Intégration d'informations sémantiques au sein de SRI

En extension de requêtes, l'ajout de connaissances sémantiques permet de préciser la requête de l'utilisateur, en ciblant plus précisément le sens de ses constituants initiaux et les rendant par conséquent moins ambigus. Lors de la phase d'indexation, les informations sémantiques visent à offrir une représentation plus riche des contenus textuels,

[11] Relation partie/tout, *e.g.* *main* et *bras*.

basée non plus sur de simples mots mais sur leur sens, favorisant ainsi un appariement plus fin entre les documents et requêtes.

Nous nous intéressons aux résultats des diverses expériences qui exploitent ces connaissances sémantiques en RI, soit pour enrichir les requêtes, soit lors de l'indexation.

2.4.2.1 Exploitation d'informations sémantiques en extension de requêtes

L'impact de connaissances sémantiques utilisées pour enrichir les requêtes varie selon le type d'information pris en compte. Pour ce qui concerne des connaissances acquises automatiquement à l'aide de méthodes numériques, les résultats obtenus dans différentes expérimentations (Peat et Willett, 1991; Gauch *et al.*, 1999; Qiu et Frei, 1995; Jing et Croft, 1994, *inter alia*) attestent de leur intérêt, puisque la plupart aboutisse à une amélioration, parfois légère, mais constante, des performances. L'approche traditionnellement adoptée consiste à repérer dans les documents les mots cooccurrant fréquemment avec les termes de la requête, et à les utiliser pour son enrichissement[12]. Cette méthode est efficace à condition que les termes utilisés pour cet enrichissement soit effectivement reliés sémantiquement aux termes originaux. Pour contrôler ce lien, plusieurs stratégies ont été proposées. Qiu et Frei (1995) par exemple prennent en considération non plus chaque terme de manière isolée mais la requête dans sa globalité. Les mots ajoutés doivent donc être proches de l'ensemble des mots de la requête et non plus d'un terme en particulier. Cette approche conduit à une amélioration significative de l'efficacité des systèmes. La particularité de l'approche de Jing et Croft (1994) est, quant à elle, de prendre en compte comme cooccurrents des termes de la requête non plus uniquement des termes simples mais également des syntagmes (nominaux, verbaux...). Là encore, les améliorations obtenues sont significatives.

Des informations sémantiques extraites par des techniques symboliques et obéissant à des contraintes linguistiques plus fortes contribuent également à l'amélioration des performances des SRI, comme l'attestent notamment les travaux présentés dans (Claveau et Sébillot, 2004b; Grefenstette, 1997) qui proposent d'enrichir les noms contenus dans une requête à l'aide de verbes qui sont liés à eux par une relation spécifique, ou ceux de (Khoo, 1995) qui utilisent pour l'expansion, des unités lexicales unies par un lien de causalité à un ou plusieurs constituants de la requête.

Enfin, une autre approche consiste à utiliser des informations sémantiques issues de bases lexicales existantes ; un nombre important de travaux (Voorhees, 1998; Mihalcean et Moldovan, 2000; Richardson *et al.*, 1996, *inter alia*) se sont ainsi appuyés sur les connaissances contenues dans la base WORDNET. Leur exploitation en extension de requêtes consiste généralement à inclure des mots qui sont sémantiquement reliés aux concepts de la requête originale en suivant les relations de synonymie ou d'hyperonymie qui les structurent très fréquemment. Chaque mot dans cette base étant associé à un certain nombre de sens différents (WORDNET répertorie par exemple pour le nom *break* 63 sens), le problème réside dans la sélection des « bons » termes (*e.g.* les synonymes)

[12]Par exemple, si les termes *document*, *requête* et *indexation* apparaissent souvent ensemble (*i.e.* de façon statistiquement significative) dans les documents de la collection, il est alors possible d'ajouter les termes *requête* et *document* à la requête *indexation*.

à utiliser pour l'extension (*i.e.* comment identifier de manière automatique que ces termes possèdent le même sens que ceux de la requête originale). Les expériences réalisées montrent que si on étend manuellement les requêtes à l'aide des informations de WORDNET, les performances des SRI peuvent être améliorées. La tentative d'automatisation de cette tâche d'extension conduit cependant à une dégradation importante des résultats. L'utilisation de ces connaissances doit donc être nécessairement associée à un traitement de désambiguïsation efficace.

D'une manière générale, il ressort de ces différentes expérimentations que l'enrichissement des requêtes à l'aide de connaissances sémantiques (quelle que soit leur provenance) est intéressant uniquement si les mots ajoutés sont véritablement liés à ceux de la requête et dépend donc de la reconnaissance de leur sens. Compte tenu de ces observations, d'autres stratégies d'expansion ont été proposées. L'une d'elles consiste à recourir à la rétroaction de pertinence (*relevance feedback*). Xu et Croft (1998), par exemple, utilisent, pour enrichir la requête, des termes issus des documents considérés comme pertinents (obtenus après un premier passage du SRI) qui cooccurrent avec tous les constituants de la requête initiale. Les résultats obtenus montrent l'efficacité de la méthode proposée (hausse de la précision moyenne comprise entre 14 et 25% selon la longueur de la requête). Mandala *et al.* (1998) proposent, quant à eux, une technique d'extension de requêtes basée sur la combinaison de connaissances sémantiques extraites à l'aide de plusieurs méthodes. L'intérêt de ce couplage est de combler les limites respectives des diverses approches. Les requêtes sont donc enrichies à l'aide d'informations extraites de WORDNET, de relations tête+modifieur obtenues par une approche symbolique, et de connaissances sémantiques acquises par repérage de cooccurrences. Une amélioration significative des performances est obtenue à condition que les pondérations des différentes informations soient bien adaptées à leurs spécificités. Le couplage de ces connaissances présente l'intérêt de désambiguïser une partie des termes de la requête.

L'apport des informations sémantiques en RI s'étant révélé, au travers des différentes expériences présentées, assez concluant lorsque celles-ci sont exploitées pour étendre les requêtes, nous nous intéressons à présent à leur efficacité lorsqu'elles sont utilisées pour enrichir les représentations des documents et requêtes lors de l'indexation.

2.4.2.2 Exploitation d'informations sémantiques pour l'indexation

Plusieurs alternatives ont été proposées pour contourner les limites des représentations à base de mots-clés traditionnellement utilisées en RI. Certains systèmes s'orientent vers une indexation dite « conceptuelle » où le concept se substitue au mot comme unité d'indexation du document ou de la requête. L'appariement document-requête correspond alors à une comparaison de concepts. D'autres optent pour une indexation dite « sémantique », qui utilise les informations sémantiques évoquées précédemment pour enrichir les représentations des documents et requêtes. Nous présentons successivement ces deux approches.

Indexation conceptuelle

Pour arriver à identifier les concepts des documents et requêtes et les mettre en correspondance, l'indexation conceptuelle s'appuie sur l'utilisation d'ontologies et utilise un formalisme de représentation des connaissances[13]. De ce fait, elle ne s'applique principalement que sur des domaines spécialisés. De manière plus précise, le principe de fonctionnement des systèmes reposant sur ce type d'indexation est le suivant : à l'aide de bases de connaissances lexicales du domaine — contenant des unités lexicales et des informations (morphologiques, sémantiques...) sur leurs propriétés — structurées par diverses relations[14], de techniques de TAL, voire d'apprentissage pour acquérir des connaissances nouvelles sur le domaine, les termes significatifs et les liens qui les unissent sont extraits des documents et requêtes, puis sont décrits à l'aide d'un formalisme de représentation de connaissances (graphes conceptuels, réseaux sémantiques, logique de description...) qui permet d'identifier les concepts et d'expliciter leurs relations sémantiques. Plusieurs travaux se sont ainsi intéressés à l'indexation conceptuelle basée sur une organisation taxonomique de la connaissance. Ils diffèrent tout d'abord par le type de formalisme utilisé pour représenter les connaissances : les expériences de (Guarino *et al.*, 1999; Chevallet, 1992; Zweigenbaum et Menelas, 1994, *inter alia*) s'appuient par exemple sur les graphes conceptuels de Sowa (1984) (l'appariement requête-documents est donc un appariement de graphes), ceux de Berrut (1990) sur un réseau sémantique inspiré des dépendances conceptuelles de Schank (1972)... Ils se distinguent également par le domaine de connaissance sur lequel ils s'appliquent : sport (Khan, 2000), médecine (Müller *et al.*, 2004; Zweigenbaum et Menelas, 1994; Berrut, 1990), informatique (Chevallet, 1992)...

D'une manière générale, l'indexation conceptuelle peut représenter une solution à certains problèmes du langage naturel, tels que la polysémie par exemple. Certains des résultats obtenus dans les expérimentations de ce type attestent de son utilité en RI, les travaux de Woods et Ambroziak (1998) obtenant par exemple une amélioration de précision et de rappel par rapport à un SRI classique. Elle nécessite cependant des ressources considérables, et fait appel à des techniques complexes et souvent coûteuses. L'indexation sémantique, plus souple et non limitée à un domaine lui est donc souvent préférée.

Indexation sémantique

Les SRI basés sur ce type d'indexation s'appuient soit sur des informations sémantiques issues de ressources construites *a priori*, soit sur des connaissances acquises sur la collection de textes, pour enrichir la représentation des documents et requêtes à l'aide de mots sémantiquement proches. Pour les premiers, l'indexation sémantique consiste, après avoir appliqué un traitement de désambiguïsation aux termes des documents et requêtes, à enrichir ceux-ci à l'aide de mots sémantiquement

[13] Pour une présentation détaillée de l'indexation conceptuelle, se référer à (Baziz, 2005).

[14] *E.g.* la relation *est-un* (*is-a* en anglais).

proches présents dans une base lexicale (prenant généralement la forme d'un ensemble de synonymes appelé *synset* dans WORDNET). Plusieurs expériences témoignent de l'intérêt du couplage « indexation à base de mots-clés – indexation à base de *synsets* ». (Smeaton *et al.*, 1995; Gonzalo *et al.*, 1998; Mihalcean et Moldovan, 2000) par exemple montrent qu'une indexation fondée sur les *synsets* conduit à une hausse de 29% de la précision, la désambiguïsation des termes étant cependant effectuée manuellement (ou semi-automatiquement pour Mihalcean et Moldovan (2000)). Globalement, une amélioration des performances des SRI est constatée lorsque les deux types d'indexation (classique et à base sémantique) sont combinés. Ces résultats sont néanmoins fortement dépendants de la qualité du traitement de désambiguïsation effectué, un module de désambiguïsation entièrement automatique pouvant conduire à une dégradation importante des performances (Voorhees, 1998).

Les SRI à indexation sémantique basée sur des connaissances sémantiques acquises automatiquement sur la collection de textes s'appuient généralement sur des informations de cooccurrence et de similarité de cooccurrents pour dériver le sens des termes et aboutir à un appariement fondé sur le sens des mots et non plus sur les seuls mots. Dans (Schütze et Pedersen, 1995) par exemple, les documents sont classés en fonction du nombre de sens qu'ils partagent avec la requête. Pour parvenir à ce résultat, chaque terme (des documents et requêtes) est associé à un sens obtenu en s'appuyant sur l'exploration de son contexte et sur le principe que les occurrences d'un mot utilisées dans le même sens partagent les mêmes contextes[15]. Les résultats obtenus (amélioration de la précision de 11%) attestent de l'intérêt de combiner une indexation sémantique à une indexation classique. Le principal avantage de cette méthode est d'être entièrement automatique. Les travaux de Rajman *et al.* (2000) proposent également d'améliorer les performances d'un SRI (vectoriel) en introduisant des connaissances sémantiques, qui sont des informations de cooccurrence d'unités linguistiques (noms, verbes...) des documents avec les termes d'indexation retenus[16]. Là encore, une amélioration significative des performances (pour un faible rappel néanmoins) est constatée.

L'indexation sémantique, quel que soit le type d'information sur lequel elle s'appuie, est, nous l'avons vu, fortement tributaire d'une désambiguïsation efficace des termes qui doit forcément dans un cadre de RI, être automatique. Nous terminons en présentant les principales méthodes de désambiguïsation automatique appliquées en RI.

2.4.3 Désambiguïsation automatique en RI

La désambiguïsation automatique est un domaine de recherche à part entière[17]. Dans un cadre de RI, de nombreux travaux de recherche se sont intéressés à ce problème central (Sanderson, 1997).

L'objectif d'un traitement de désambiguïsation est de distinguer les différents sens que peut porter un terme. En RI, on compte principalement deux approches pour parve-

[15] La méthode proposée est décrite plus précisément dans (Schütze et Pedersen, 1995).

[16] Pour un détail de cette approche, *cf.* (Rajman *et al.*, 2000; Besançon, 2002).

[17] Pour un état de l'art plus complet des différentes méthodes proposées, se référer à (Kilgarriff et Palmer, 2000; Audibert, 2003, *inter alia*).

nir à ce but. La première est basée sur une connaissance *a priori* du nombre de sens d'un mot et repose généralement sur des données lexicales existantes telles que des dictionnaires, des thésaurus ou des bases lexicales. Il s'agit alors de reconnaître le sens d'une occurrence donnée d'un mot parmi l'ensemble de ses sens répertoriés. Cette approche a donné lieu à un nombre important d'expérimentations, essentiellement avec WORDNET. Généralement, la méthode utilisée (Towell et Voorhees, 1998; Uzuner *et al.*, 1999; Mihalcean et Moldovan, 2000) consiste à examiner le contexte d'utilisation d'un mot ambigu pour tenter de déterminer l'ensemble des synonymes adéquats pour une occurrence donnée et donc son sens. L'idée-clé est que les mots qui sont utilisés avec le même contexte ont des sens similaires ou reliés. Les informations contenues dans WORDNET couplées aux unités lexicales du contexte permettent alors d'identifier le *synset* du mot ambigu dans son contexte. Les expériences fondées sur cette méthode de désambiguïsation révèlent cependant, nous l'avons dit précédemment, des résultats globalement mitigés ne présentant que pas ou peu de hausse des performances des SRI. Ces résultats peuvent s'expliquer par la difficulté à désambiguïser les mots contenus dans les requêtes courtes (composées d'un ou deux mots), par la couverture insuffisante des ressources utilisées, par leur inadéquation à être utilisées sur des collection de textes particulières, voire par la granularité trop fine des sens qu'elles répertorient, difficilement compatibles avec la RI. La seconde approche pour procéder à la désambiguïsation des termes en RI s'appuie sur des informations extraites directement des documents concernés, essentiellement à l'aide de méthodes numériques. La désambiguïsation consiste dans ce cas à déterminer si un mot ambigu donné est utilisé avec le même sens dans deux occurrences distinctes, plutôt que de chercher à associer un sens précis à ce mot. En s'appuyant uniquement sur le contexte des mots pour leur désambiguïsation, ce type d'approche ne nécessite par conséquent aucune connaissance *a priori* sur le nombre de sens possibles d'un mot. Les expériences de (Schütze et Pedersen, 1995; Rajman *et al.*, 2000) (décrites en section 2.4.2.2) vont dans ce sens. Schütze et Pedersen (1995) en particulier propose de regrouper les mots selon les similitudes de leurs contextes, chaque groupe étant supposé représenter un sens. Les sens déterminés par cette méthode sont liés aux documents sur lesquels ils ont été appris et sont donc peu réutilisables.

Les résultats de ces différentes expériences ont mis en valeur la difficulté de disposer d'outils de désambiguïsation efficaces en RI. Les méthodes basées sur des ressources pré-construites sont notamment limitées, nous l'avons vu, par le caractère statique des informations qu'elles contiennent et qui ne sont pas nécessairement adaptées à une collection donnée. Les connaissances acquises en corpus apparaissent potentiellement plus prometteuses ; les résultats observés suite à leur intégration en RI restent cependant insuffisants.

2.4.4 Bilan de l'apport de connaissances sémantiques en RI

Au travers de ces différentes expérimentations, l'apport effectif de l'exploitation de connaissances sémantiques en RI est difficile à déterminer. Il dépend d'un nombre important de facteurs. L'un d'entre eux est la provenance ou mode d'acquisition de ces informations. Les connaissances issues de ressources généralistes construites *a priori* sont

principalement limitées par le fait qu'elles ne sont pas toujours assez spécifiques, et il est souvent difficile d'identifier le sens d'un mot, notamment dans des documents spécialisés. Les informations extraites à l'aide de méthodes numériques ne prennent généralement pas en compte les cas où deux mots peuvent être similaires sans apparaître nécessairement au sein des mêmes documents. L'apport de ces informations est fortement lié aux collections de documents et requêtes utilisées. Les connaissances acquises à l'aide de techniques symboliques sont encore peu exploitées en RI ; elles sont plus fiables mais leur méthode d'extraction souvent moins portable. Enfin, le choix d'utiliser ces connaissances sémantiques pour étendre les requêtes ou pour enrichir les représentations des documents et requête reste difficile. Comme pour les informations syntaxiques, leur exploitation lors de l'indexation conduit nécessairement à une adaptation des modèles de RI, peu propices à les accueillir.

2.5 Vers un autre couplage TAL-RI

Ce chapitre a proposé un tour d'horizon des principales contributions actuelles des techniques du TAL à la RI. Le but recherché était de montrer qu'une analyse plus fine des documents et requêtes permettait d'extraire et d'intégrer au sein des SRI des informations plus riches que de simples mots-clés.

À travers cet état de l'art, nous avons pu constater que le couplage TAL-RI a fait l'objet d'un nombre important de travaux de recherche. Une grande variété d'informations linguistiques particulières, appartenant aux niveaux morphologique, syntaxique et sémantique de la langue a en effet été exploitée en RI, et l'apport respectif de ces diverses connaissances sur les performances des systèmes a été évalué. L'impact des méthodes utilisées pour leur acquisition a également été étudié, différenciant les informations dérivées directement des documents à l'aide de techniques numériques ou symboliques de celles issues de ressources pré-existantes. Plusieurs modes d'intégration de ces connaissances au sein des SRI ont aussi été expérimentés, soit lors de l'indexation des documents et requêtes, soit directement lors de la recherche en étendant les requêtes, voire même en post-traitement des systèmes.

Malgré le nombre et la diversité des pistes explorées, il reste difficile de dresser un bilan précis de l'apport de ces informations linguistiques en RI. Les résultats obtenus dans les expérimentations proposées sont, pour la plupart, nous l'avons vu, contradictoires, ou plus exactement dépendants de multiples paramètres.

À partir de ce constat, il semble donc nécessaire d'explorer de nouveaux axes de recherche pour tenter d'obtenir des résultats plus tranchés sur l'intérêt du TAL pour la RI. Une piste possible serait de chercher à évaluer l'apport d'autres informations linguistiques non encore exploitées en RI. Sur le plan syntaxique par exemple, d'autres connaissances structurelles et structurantes telles que les entités nommées (noms de lieux, de personnes...) ou des informations de positionnement des mots (distance, ordre...) sont extractibles. Sur le plan sémantique, rares également sont les expériences qui s'intéressent à la notion de thème, c'est-à-dire qui cherchent à identifier le ou les thèmes

des documents et à établir une correspondance avec la thématique des informations recherchées par l'utilisateur.

Nous faisons le choix dans le cadre de cette thèse de ne pas expérimenter la prise en compte de nouvelles connaissances linguistiques[18], mais de chercher plutôt, en s'appuyant sur les diverses observations qui ont pu être faites au cours de cet état de l'art, à proposer une nouvelle façon d'explorer le couplage TAL-RI en exploitant *autrement* les diverses connaissances existantes.

Pour cela, nous nous appuyons plus précisément sur le constat que la plupart des expérimentations déjà réalisées cherchent généralement à évaluer l'apport d'une connaissance linguistique particulière au sein d'un SRI. L'information qui est alors prise en compte appartient à l'un des trois niveaux de langue suivants : le niveau morphologique, syntaxique ou sémantique. Partant du principe, déjà évoqué en introduction générale que, dans la langue, ces trois niveaux sont dépendants les uns des autres, il nous semble intéressant, dans le but d'obtenir une caractérisation plus riche des documents et requêtes et d'exploiter par conséquent pleinement toute la richesse de la langue, de combiner et d'intégrer au cœur d'un même système, des informations linguistiques appartenant à ces trois niveaux de langue.

L'intégration d'informations linguistiques multi-niveaux ne peut se limiter néanmoins à une simple méthode de multiplication de connaissances au sein d'un SRI. En effet, étant donné la nature différente des informations prises en compte et compte tenu du fait que l'on connaît leur comportement en RI uniquement lorsqu'elles sont exploitées individuellement, plusieurs questions fondamentales doivent au préalable être posées. Il s'agit tout d'abord de s'interroger sur l'efficacité respective de ces informations linguistiques pour retrouver des documents pertinents. Elles n'ont pas toutes en effet le même impact sur les performances des SRI et devront par conséquent être considérées différemment au sein du couplage. De plus, il est nécessaire d'examiner la façon dont elles se comportent les unes par rapport aux autres. Certaines informations peuvent avoir un impact similaire sur les performances des SRI (*e.g.* retrouver les mêmes documents pertinents). Dans ce cas, il s'avérerait inutile de multiplier des connaissances linguistiques dont l'apport en RI est identique. D'autres, au contraire, peuvent intervenir de manière complémentaire, leur couplage permettant alors de retrouver plus de documents pertinents.

Le chapitre suivant propose donc d'apporter des éléments de réponse à l'ensemble de ces questions. Nous cherchons plus précisément à évaluer l'intérêt de coupler des informations linguistiques multi-niveaux en RI en nous focalisant d'une part sur l'apport respectif des diverses connaissances prises en compte et, d'autre part, sur les relations qu'elles entretiennent lorsqu'elles sont combinées. Pour cela, nous proposons, à partir d'une plate-forme réalisée pour intégrer en parallèle ces connaissances multi-niveaux au sein d'un même système, une analyse originale des corrélations entre ces diverses informations du point de vue de leur efficacité à retrouver des documents pertinents.

[18] Ce choix est motivé notamment par le fait que nous souhaitons éviter d'être confrontée aux mêmes problèmes que les expériences présentées ici, à savoir l'obtention de résultats variables car tributaires d'un nombre important de facteurs.

Chapitre 3

Pertinence du couplage d'informations linguistiques multi-niveaux en RI

Résumé : Dans l'optique d'étudier le couplage TAL-RI sous un nouvel angle, ce chapitre propose d'évaluer l'intérêt de combiner en RI des informations linguistiques multi-niveaux (*i.e.* appartenant à la fois aux niveaux morphologique, syntaxique et sémantique de la langue). L'objectif ici est de déterminer si ces connaissances de natures variées, lorsqu'elles sont intégrées de manière simultanée au sein d'un SRI, ont toutes le même impact sur les performances des systèmes, et si elles entretiennent des relations entre elles, c'est-à-dire si elles sont redondantes ou au contraire complémentaires pour retrouver des documents pertinents.

Mots-clés : couplage d'informations linguistiques multi-niveaux, intégration au sein d'un SRI, analyse de corrélations.

3.1 Introduction

L'exploitation de techniques du TAL en RI se résume essentiellement, nous l'avons vu au chapitre précédent, à la prise en compte et à l'intégration, au sein des systèmes, d'un type particulier d'information linguistique mono-niveau (*i.e.* d'ordre morphologique, syntaxique ou sémantique). Bien que ce genre d'approche mette en évidence l'intérêt de recourir à certaines de ces connaissances pour améliorer les performances des SRI, il n'offre qu'une évaluation partielle de l'apport du TAL en RI. Dans le but d'aborder le couplage TAL-RI sous un nouvel angle, nous proposons ici des pistes de recherche qui visent à exploiter différemment les informations linguistiques au sein des systèmes. Nous cherchons plus précisément à évaluer l'intérêt en RI de coupler et d'intégrer ensemble au cœur d'un même système des connaissances appartenant à tous les niveaux de la langue. L'hypothèse qui est faite est que la combinaison de ces informations multi-niveaux, plus apte à refléter la richesse de la langue, doit offrir une meilleure caractérisation des

contenus textuels et, par conséquent, contribuer à améliorer les performances des SRI. Selon cette approche, chaque document (ou requête) n'est plus associé à un seul type de descripteur mais à une combinaison de diverses représentations (chacune correspondant à un type d'information linguistique particulier).

Le couplage d'informations linguistiques multi-niveaux ne peut toutefois se résumer à l'intégration d'une multitude de descripteurs différents au sein d'un même système. En effet, étant donné la diversité des connaissances prises en compte et compte tenu du fait que l'on connaît leur apport et leur comportement en RI uniquement à partir de leur exploitation individuelle, il nous semble primordial d'étudier la pertinence de cette combinaison en nous posant un certain nombre de questions fondamentales et en proposant des moyens adéquats pour y répondre : à quels résultats peut-on s'attendre ? Les traitements effectués sont-ils complémentaires ou concurrents ? Les éventuels gains de performances sont-ils susceptibles de s'additionner ? Les études menées dans ce chapitre visent par conséquent deux objectifs. Le premier est de mesurer, dans un cadre homogène et sur des données identiques, l'efficacité des divers types d'informations linguistiques pour retrouver des documents pertinents. Le second est d'obtenir une réponse quant à l'intérêt ou non d'intégrer au sein d'un même SRI de manière conjointe des connaissances linguistiques de plusieurs niveaux de langue. Pour ce faire, nous proposons, à partir d'une plate-forme réalisée pour intégrer en parallèle dans un SRI des connaissances multi-niveaux, une analyse originale des corrélations entre ces diverses informations du point de vue de leur efficacité en RI à retrouver les documents pertinents.

Après avoir positionné notre approche par rapport aux travaux existants (section 3.2), nous présentons (section 3.3) les informations linguistiques multi-niveaux que nous exploitons et l'architecture de test bâtie pour les combiner et les intégrer au sein d'un SRI. Nous nous intéressons ensuite (section 3.4) d'une part à l'évaluation, dans un cadre unifié, de l'impact individuel de ces diverses connaissances et, d'autre part, à l'analyse des relations susceptibles d'exister entre elles. Ces différentes études nous permettent de dresser enfin (section 3.5) un bilan de l'intérêt du couplage de connaissances morphologiques, syntaxiques et sémantiques au sein d'un SRI.

3.2 Travaux sur l'exploitation d'informations linguistiques multi-niveaux en RI

Peu de travaux à notre connaissance proposent d'intégrer des informations linguistiques multi-niveaux au sein d'un même SRI. Comme nous l'avons vu au chapitre précédent, la plupart cherche généralement à évaluer l'apport d'un type d'information particulier d'ordre morphologique, syntaxique ou sémantique. Sans aller jusqu'à exploiter de manière simultanée les trois niveaux, quelques-uns s'intéressent toutefois à la combinaison de connaissances mono-niveau, voire bi-niveaux, pour enrichir les représentations des documents et requêtes. Pour le couplage d'informations appartenant à un même niveau de langue, nous pouvons par exemple citer les travaux de Zhai *et al.* (1997) qui intègrent en parallèle au sein d'un même SRI plusieurs représentations différentes d'un

même document, chacune correspondant à un type de structure syntaxique particulier qu'ils souhaitent évaluer (*cf.* section 2.3.2 du chapitre précédent). Pour la combinaison d'informations bi-niveaux, plusieurs stratégies sont généralement adoptées. Une première vise à représenter les documents en s'appuyant sur le couplage d'informations appartenant à l'un des trois niveaux de la langue avec des termes simples. Il s'agit par exemple des expériences qui prennent en compte à la fois des termes simples et complexes (ou des syntagmes) au sein d'un même SRI (Haddad, 2002; Pedersen et Bruce, 1997; Arampatzis *et al.*, 1996, *inter alia*) ou celles qui proposent une indexation basée sur la combinaison termes simples et *synsets* (Gonzalo *et al.*, 1998). Une seconde stratégie consiste à mêler des connaissances de deux niveaux de langue différents, *e.g.* des niveaux morphologique et sémantique. Les documents sont alors par exemple représentés par la combinaison de lemmes et d'entités nommées pour (Friburger et Maurel, 2002) ou de *stems* et d'informations sémantiques issues de WORDNET pour (Voorhees, 1999). D'autres expérimentations cherchent également à prendre en compte diverses connaissances linguistiques (là encore généralement bi-niveaux) non pas pour enrichir les représentations des documents mais pour étendre les requêtes (Mandala *et al.*, 1998).

Quelques rares études proposent toutefois d'exploiter pleinement la richesse de la langue pour la représentation des contenus textuels en prenant en compte de manière simultanée les trois niveaux de la langue. Strzalkowski *et al.* (Strzalkowski *et al.*, 1999; Perez-Carballo et Strzalkowski, 2000) proposent ainsi un système de descripteurs en parallèle, construits pour une collection donnée, où chaque descripteur reflète une stratégie particulière de représentation linguistique des textes. Dans ce système, la première stratégie retenue vise à représenter les documents et requêtes du point de vue des informations morphologiques qu'ils contiennent. Le premier type de descripteur considéré correspond donc aux racines de l'ensemble des termes de la collection. La deuxième stratégie propose de représenter les textes à l'aide de connaissances d'ordre syntaxique, qui se manifestent sous la forme de syntagmes normalisés en tête+modifieur acquis par le biais d'une analyse syntaxique des documents et requêtes. Ces syntagmes constituent le deuxième type de descripteur utilisé. Enfin, la dernière sorte de représentation exploite des informations d'ordre sémantique. Le troisième descripteur correspond alors à l'ensemble des entités nommées extraites de la collection. Ces différentes représentations sont ensuite utilisées par le processus de recherche pour apparier documents et requêtes. Le classement final des documents est obtenu en fusionnant les résultats individuels de chaque descripteur. Enfin, les performances de ce « méta-système » de recherche peuvent être optimisées en utilisant les meilleurs résultats de chaque représentation. Les expériences menées avec cette architecture montrent que les informations linguistiques peuvent contribuer à améliorer les résultats des SRI, mais de manière très modeste.

Le système que nous proposons pour l'évaluation de l'intégration de connaissances multi-niveaux s'inspire de ce dernier travail puisque nous ré-utilisons l'idée de cette architecture qui nous paraît pertinente pour représenter de manière simultanée plusieurs informations linguistiques sous la forme de descripteurs mis en parallèle. Notre approche s'en distingue toutefois d'une part par la nature et la diversité des informations linguistiques prises en compte et, d'autre part, par sa finalité. Notre objectif n'est pas, comme

nous l'avons déjà évoqué, de concevoir un système qui incorpore une multitude de descripteurs mais d'étudier comment ces derniers se comportent les uns par rapport aux autres dans la recherche de documents pertinents et d'évaluer la pertinence d'un tel couplage.

Nous présentons dans la section suivante les informations linguistiques multi-niveaux que nous avons choisi d'exploiter et l'architecture de test mise en place pour leur intégration en parallèle au sein d'un même système.

3.3 Architecture pour le couplage d'informations linguistiques multi-niveaux en RI

Comme nous l'avons déjà souligné, notre but n'est pas de chercher à évaluer l'apport en RI de nouvelles connaissances linguistiques particulières, mais d'étudier l'intérêt de combiner des informations de natures variées, ayant déjà été exploitées individuellement en RI. Nous revenons plus précisément dans un premier temps (section 3.3.1) sur les connaissances d'ordre morphologique, syntaxique et sémantique que nous avons retenues pour représenter les documents et requêtes, puis décrivons (section 3.3.2) l'architecture envisagée pour leur combinaison et leur intégration au sein d'un même SRI.

3.3.1 Informations linguistiques multi-niveaux

Pour sélectionner les informations linguistiques qui seront ensuite combinées pour enrichir la représentation textuelle des documents et requêtes, nous nous sommes appuyée sur les nombreux travaux existants, présentés au chapitre précédent, cherchant à évaluer l'apport de connaissances linguistiques mono-niveau en RI. Ces diverses études nous ont permis de recenser les informations (de nature morphologique, syntaxique et sémantique) susceptibles d'être insérées au sein d'un SRI, tout en nous donnant une idée de leur potentielle efficacité. Pour effectuer notre choix parmi les connaissances disponibles, nous nous sommes imposé deux contraintes fortes. La première concerne la volonté de recourir à des informations linguistiques « standards », *i.e.* traditionnellement exploitées en RI. Notre objectif, en effet, n'est pas d'évaluer l'impact d'une information particulière sur les performances des systèmes mais l'intérêt de combiner des connaissances habituellement prises en compte en RI. Il ne s'agit donc pas de chercher à exploiter des informations très fines nécessitant des méthodes complexes pour leur acquisition, mais plutôt d'utiliser des connaissances facilement extractibles à l'aide d'outils disponibles et communément utilisés en RI. La seconde contrainte est directement liée à la collection de documents et requêtes utilisée pour nos expérimentations. Nous plaçons nos travaux dans un cadre d'évaluation traditionnel de RI, et utilisons par conséquent des collections de référence qui nous permettent de comparer nos résultats avec ceux issus des approches classiques de RI. Les outils utilisés pour l'extraction des informations linguistiques doivent être adaptés aux spécificités de ces collections (décrites plus en détail en section 3.4.1). Ils doivent dans notre cas être capables de

manipuler des volumes de données assez importants (la collection utilisée est composée d'environ 175000 documents) en anglais.

Ces contraintes définies, nous présentons successivement les informations linguistiques de nature morphologique, syntaxique et sémantique que nous avons choisi d'exploiter et les outils et méthodes utilisées pour leur acquisition. L'ensemble des traitements évoqués ici est appliqué sur les documents et requêtes directement issus de la collection ; seul un pré-traitement supprimant certaines balises est effectué. Tous les mots des textes et questions sont pris en compte et considérés comme des termes d'indexation potentiels. Leur pondération et la suppression des mots-vides sont effectuées ultérieurement (lors de l'intégration des diverses représentations linguistiques au sein du SRI).

Informations de nature morphologique

Pour le niveau morphologique, nous avons choisi de considérer trois types de connaissances en particulier. Nous exploitons tout d'abord des informations d'ordre flexionnel. Un traitement de lemmatisation (effectué par le biais de l'outil TREETAGGER[1]) est appliqué à tous les documents et requêtes et consiste à identifier, pour chaque mot des textes et questions, son lemme, *i.e.* sa forme de base débarrassée de ses flexions. Ce traitement présente l'avantage, nous l'avons vu, de gérer une partie du problème de la variation morphologique. Nous prenons également en compte des informations morphologiques d'ordre flexionnel et dérivationnel. Pour cela, nous appliquons aux textes (et questions) une procédure de racinisation (*stemming*) qui permet d'extraire pour chaque mot sa pseudo-racine (*stem*). Après avoir évalué et comparé différents outils (notamment les *stemmers* de Porter (1980), de Lovins (1968) ou de Paice (1990)), nous nous sommes appuyée sur l'algorithme de Porter (1980), jugé le plus performant dans le cadre de notre collection, pour la normalisation des variantes morphologiques. Le choix de recourir à la fois à un traitement de lemmatisation et de racinisation est notamment motivé par le fait que les expériences existantes (*cf.* chapitre précédent) n'ont pas permis de trancher sur la supériorité d'une des deux techniques. En couplant ces deux informations au sein de notre architecture de SRI, nous pourrons d'une part évaluer celle qui a le plus d'impact sur les performances du SRI et, d'autre part, déterminer si les informations de lemmes et de racines sont redondantes ou au contraire complémentaires.

Enfin, le dernier type d'information utilisé est d'ordre morpho-syntaxique. Une analyse morpho-syntaxique des documents et requêtes est réalisée dans le but d'associer à chaque mot sa catégorie grammaticale (nom, verbe, adjectif...). Le principal intérêt de cet étiquetage est qu'il permet d'opérer un premier traitement de désambiguïsation des termes. L'étiqueteur utilisé ne peut en effet associer qu'une seule étiquette à chaque mot ; il doit par conséquent choisir parmi toutes les catégories possibles d'un mot (*e.g.* *brise* qui peut à la fois être un nom ou un verbe) celle qui correspond précisément au

[1] TREETAGGER est disponible à l'adresse suivante : `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`.

terme dans la phrase considérée, en s'appuyant notamment sur son contexte d'apparition. Après avoir comparé deux outils pour l'anglais — l'étiqueteur à base de règles de Brill (1992) et l'étiqueteur probabiliste TREETAGGER qui s'appuie sur des arbres de décision (Schmid, 1997) — nous avons sélectionné TREETAGGER, évalué comme le plus efficace sur notre collection de documents.

Informations de nature syntaxique

Pour le niveau syntaxique de la langue, nous avons retenu trois principales structures permettant de rendre compte des relations et dépendances entre les mots. Nous prenons en compte tout d'abord les termes complexes. Leur reconnaissance est effectuée à l'aide de l'outil FASTR (Jacquemin et Royauté, 1994) qui, par le biais d'une analyse morpho-syntaxique des textes et de l'utilisation de méta-règles linguistiques, permet l'extraction automatique de termes complexes mais également la normalisation de leurs variantes (les termes *blood flow* et *flow of blood* par exemple sont considérés comme équivalents). Nous exploitons également des informations de bigrammes et de trigrammes. L'outil utilisé (NGRAM STATISTIC PACKAGE (NSP)), basé sur une approche essentiellement numérique, consiste à repérer et à extraire les suites de n termes (dans notre cas n correspond à 2 (bigrammes) et à 3 (trigrammes)) qui apparaissent de manière statistiquement significative dans la collection[2]. Des mots qui apparaissent fréquemment ensemble peuvent être, comme nous l'avons vu au chapitre précédent, pertinents pour représenter les documents et requêtes. Nous extrayons séparément les bigrammes et les trigrammes, chacun représentant une information linguistique particulière. Nous proposons enfin d'exploiter les syntagmes nominaux présents dans les textes et les questions. Nous utilisons pour cela l'outil développé par Ramshaw et Marcus (1995) qui, par le biais d'une méthode combinant à la fois des techniques numérique et symbolique d'acquisition[3], identifie au sein des textes les syntagmes nominaux.

Ces connaissances sont donc extraites automatiquement à partir des documents et requêtes à l'aide de méthodes différentes, symboliques pour les termes complexes, numériques pour les bigrammes et trigrammes, et mixtes pour les syntagmes nominaux.

Informations de nature sémantique

Les informations sémantiques utilisées sont issues pour la plupart d'une ressource généraliste pré-existante : WORDNET. Comme évoqué en section 2.4.3, il est nécessaire néanmoins, si l'on souhaite exploiter ces connaissances, de procéder au préalable à un traitement de désambiguïsation des mots des documents et requêtes. Nous nous appuyons pour cela sur le module de désambiguïsation proposé par Pedersen *et al.* (2004) dont le principe général consiste à sélectionner le sens d'un terme donné,

[2]Pour le traitement des bigrammes et trigrammes, un filtre est opéré afin de ne retenir que les suites constituées de mots pleins.

[3]La technique utilisée s'appuie plus précisément sur l'utilisation d'un algorithme d'apprentissage de règles de transformation à partir de corpus étiquetés morpho-syntaxiquement.

parmi tous ses sens possibles répertoriés dans WORDNET, en fonction du sens des mots apparaissant dans son contexte. Il s'agit plus précisément de calculer une distance sémantique entre les différents sens[4] du terme à désambiguïser et les sens des mots de son contexte[5]. Ce traitement est assez basique (nous utilisons les paramètres par défaut, notamment pour ne pas augmenter les temps de traitement) ; une procédure plus complexe aurait été plus coûteuse, risquant d'aller à l'encontre des contraintes que nous sommes fixées en introduction de cette section. Les mots non désambiguïsés (correspondant généralement aux termes absents de WORDNET) sont ramenés à leur forme lemmatisée. Pour les mots qui sont associés à un sens unique, nous pouvons alors extraire les informations de WORDNET qui nous intéressent. Nous récupérons plus précisément pour chaque terme désambiguïsé :

- son étiquette sémantique, *i.e.* son numéro de sens dans WORDNET, comme par exemple *retrieve#v#2* qui correspond au deuxième sens du verbe *retrieve* dans la ressource,
- l'ensemble de ses synonymes,
- et l'ensemble des mots reliés (relations inter-catégorielles) à ce terme par un lien de morphologie dérivationnelle. Par exemple, le nom *adoption* est lié dans WORDNET au cinquième sens du verbe *adopt#v#5* qui a lui-même pour synonyme le verbe *take in*. Nous formons donc à partir de ces informations la famille de mots suivante : *adoption, adopt, take in*[6].

La dernière information sémantique prise en compte correspond aux entités nommées. Après avoir évalué plusieurs outils (notamment GATE (Cunningham, 2002)) et constaté la faible qualité des informations obtenues pour notre collection, nous utilisons une méthode plus basique mais que nous considérons comme plus efficace. Nous nous appuyons uniquement sur l'étiquetage morpho-syntaxique des textes effectué à l'aide de TREETAGGER et extrayons tous les mots associés à l'étiquette *nom propre*. Ce traitement est naïf parce qu'il ne permet notamment pas de récupérer, contrairement aux outils de reconnaissance des entités nommées, les noms propres formés de plusieurs unités (le prénom et le nom d'une personne sont considérés comme deux mots différents par exemple).

Nous venons de présenter les onze types d'informations linguistiques multi-niveaux qui ont été extraits automatiquement des documents et requêtes par le biais d'outils et techniques du TAL. Comme nous l'avons dit, nous avons fait le choix d'exploiter des connaissances facilement extractibles et manipulables dans un cadre de RI plutôt que des informations peut-être plus pertinentes d'un point de vue linguistique mais plus complexes à acquérir. Chaque type de connaissance extrait d'un document (*e.g.* l'ensemble des lemmes qu'il contient) correspond à un descripteur (ou index). À la suite de ces divers traitements linguistiques, un document (ou une requête) de la collection

[4]Dans notre cas, un sens prend la forme de définitions (*gloses* dans WORDNET).

[5]Nous utilisons plus particulièrement, pour évaluer la proximité sémantique entre deux sens, une variante de la mesure de Lesk, proposée par Banerjee et Pedersen (2003).

[6]Ces familles regroupent donc des termes liés à la fois morphologiquement et sémantiquement au mot initial. C'est pourquoi nous avons choisi de classer ces informations comme appartenant au niveau sémantique et non morphologique de la langue.

peut donc être représenté par onze descripteurs différents. Il peut être vu comme un ensemble de lemmes, de racines, de termes simples associés à leur étiquette grammaticale, de termes complexes, de bigrammes, de trigrammes, de groupes nominaux, de noms propres, de termes simples associés à leurs étiquettes sémantiques, de termes simples associés à un groupe de synonymes ou encore à un groupe de mots reliés morphologiquement. Nous proposons également de représenter les documents (et requêtes) à l'aide d'un index « standard ». Nous extrayons pour cela l'ensemble des termes simples qu'ils contiennent. Nous proposons en annexe B l'exemple d'un document représenté par ces 12 descripteurs différents (11 descripteurs linguistiques + 1 descripteur composé de termes simples). Nous décrivons à présent la façon dont ces multiples représentations peuvent être combinées et intégrées au sein d'un même SRI.

3.3.2 Intégration des informations linguistiques multi-niveaux au sein du SRI

Les mécanismes traditionnels de RI, tels qu'ils ont été présentés au chapitre 1, sont généralement conçus pour manipuler un type de représentation des documents et requêtes (un descripteur par document) à la fois. La combinaison des informations multi-niveaux décrites précédemment au sein d'une seule représentation (au sein d'un même index) n'est pas pertinente dans notre cas puisque, avec cette méthode, nous obtiendrions à la suite de la phase d'appariement une seule liste de résultats. Nous ne pourrions pas par conséquent évaluer la contribution respective de chaque information prise en compte, ni étudier les relations qu'elles entretiennent. Certains modèles de RI offrent la possibilité de combiner plusieurs représentations des documents. C'est le cas par exemple des réseaux d'inférence (*cf.* section 1.3.3) qui proposent de décrire les documents selon différents niveaux d'abstraction (Turtle et Croft, 1991). Dans le cadre du modèle vectoriel dans lequel nous nous plaçons, plusieurs solutions ont également été proposées pour contourner les limites d'une représentation unique. L'une d'elles consiste à utiliser le modèle vectoriel étendu introduit par Fox (1983) qui propose de considérer un vecteur de documents comme une combinaison de sous-vecteurs, chacun pouvant représenter un type d'information particulier. La similarité entre deux vecteurs étendus est alors calculée comme une combinaison linéaire des différents sous-vecteurs. Bien que cette approche nous permette d'intégrer simultanément nos différentes informations linguistiques, nous obtiendrions également au final avec elle une seule liste de résultats. C'est pourquoi nous choisissons ici de nous appuyer plutôt sur une architecture telle que celle proposée par Strzalkowski *et al.* (décrite en section précédente) qui nous semble plus souple et plus facilement manipulable. Il s'agit de concevoir un système de descripteurs en parallèle, où chaque index reflète une représentation linguistique particulière des documents et requêtes. La phase d'appariement consiste alors à comparer chacun des descripteurs des documents au même type de descripteur associé à la requête. Ce mécanisme de mise en correspondance des différentes représentations permet d'associer à chaque couple document-requête un score de pertinence et d'obtenir, pour chaque représentation, une liste des documents classés par ordre de pertinence par rapport aux requêtes. Dans les travaux de Strzalkowski *et al.*, ces différentes listes sont fusionnées

pour obtenir un classement unique des documents. Dans nos travaux, nous exploitons séparément les listes obtenues à la suite de l'appariement de chacun des descripteurs.

De manière plus précise, l'architecture proposée peut être synthétisée de la façon suivante : les documents et requêtes passent tout d'abord par un module d'analyse linguistique (combinant les différents outils décrits précédemment) qui permet d'obtenir 12 représentations différentes d'un même document (ou requête), comme illustré sur la figure 3.1.

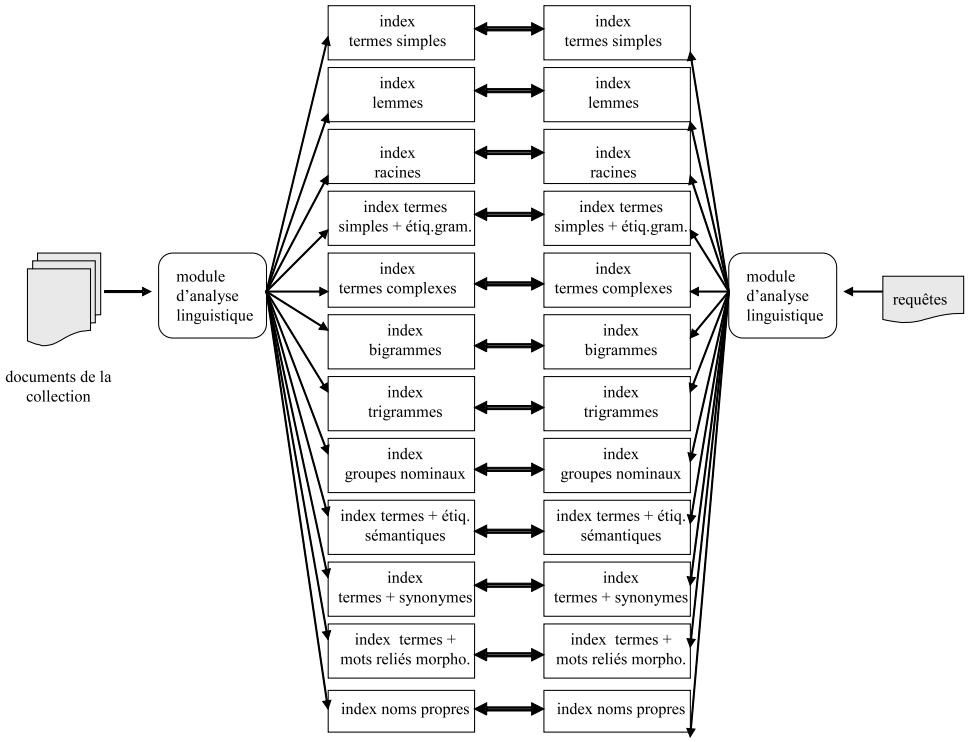

FIG. 3.1 – Représentation multi-index des documents et requêtes

Ces représentations, après avoir été au préalable formatées (dans notre cas, il s'agit de les transformer au format SGML imposé par le SRI utilisé), sont ensuite intégrées de manière parallèle au sein du SRI. Le SRI utilisé est LEMUR[7]. Il compare (pour les 12 index pris en compte) chacune des représentations des documents à celle correspondante des requêtes et calcule un score de similarité (nous utilisons la mesure probabiliste *BM25* d'Okapi qui, après plusieurs expérimentations, a été jugée comme la plus performante). Cette phase d'appariement nous permet d'obtenir pour chacun des index une liste de documents classés par ordre de pertinence décroissante par rapport aux requêtes. Ce processus est présenté en figure 3.2. Nous obtenons finalement 12 listes ordonnées de résultats.

[7]Cet outil est disponible à l'adresse suivante : `http://www.lemurproject.org/`.

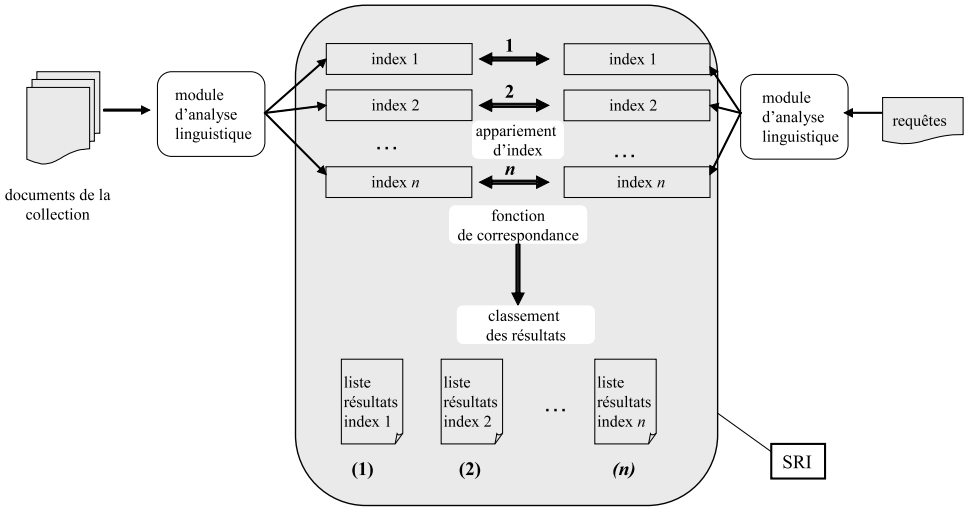

FIG. 3.2 – Intégration au sein du SRI des représentations multi-index

Pour intégrer ces différents index, nous avons dû procéder à quelques adaptations du SRI. Nous avons notamment dû modifier la définition de l'unité d'indexation initialement proposée considérant un terme d'indexation comme un terme simple, *i.e.* une chaîne de caractères comprise entre deux délimiteurs (l'espace). Par le biais des informations linguistiques prises en compte, nous exploitons en effet des unités d'indexation plus complexes, *e.g.* des suites d'unités lexicales (termes complexes, groupes nominaux, trigrammes, bigrammes), des structures « termes + étiquettes » (étiquettes grammaticales ou sémantiques) ou encore des structures de type « termes + ensemble de mots » (synonymes, mots reliés).

Cette architecture présente donc l'avantage d'intégrer en parallèle les différentes représentations linguistiques des documents et requêtes au sein du SRI. Elle nous permet d'obtenir, pour chacun des index intégrés, la liste ordonnée des documents retournés par le SRI à la suite de leur appariement. Ce sont sur ces listes, et plus précisément sur les rangs assignés aux documents, que nous allons nous appuyer pour étudier l'apport respectif de chacune des informations linguistiques prises en compte et les relations qu'elles entretiennent. Cette étude fait l'objet de la section suivante.

3.4 Informations linguistiques : intérêt individuel et pertinence du couplage

Pour évaluer l'intérêt de coupler diverses informations linguistiques au sein d'un SRI, il est, dans un premier temps, nécessaire d'avoir une idée précise de l'impact respectif de chacune de ces connaissances prise en compte indépendamment sur les performances du système. C'est ce que nous regardons en section 3.4.2. Nous nous intéressons ensuite plus précisément à l'étude des liens susceptibles d'exister entre ces informations, en cher-

chant à établir si ces dernières sont complémentaires ou au contraire redondantes. Pour cela, nous procédons (section 3.4.3) à une analyse approfondie des corrélations entre les différentes listes de documents retournées par le SRI. Pour mettre en perspective la façon dont elles sont liées les unes aux autres, nous proposons enfin d'établir (section 3.4.4) des classes d'informations linguistiques construites en fonction de leur comportement sur les performances. Nous présentons au préalable (section 3.4.1) la collection de test sur laquelle nous nous appuyons pour effectuer nos expérimentations.

3.4.1 Collection de test

Les documents et les requêtes sur lesquels nous nous appuyons pour nos évaluations proviennent de la collection TIPSTER utilisée lors des campagnes d'évaluation TREC. Cette collection contient des documents provenant de différentes sources (articles de journaux, dépêches de presse, notices de brevets...) couvrant une grande variété de domaines (politique, économie, informatique, énergie...), des requêtes et des jugements de pertinence. Pour nos expérimentations, nous utilisons plus particulièrement une sous-partie de ce corpus qui regroupe environ 175000 articles de journaux issus du *Wall Street Journal* des années 1986 à 1992. La particularité de cette sous-collection est de couvrir des thématiques variées (politique, économie..., *i.e.* tous les domaines liés à l'actualité), et de regrouper des documents (chacun correspondant à un article et étant associé à un identifiant unique) de longueurs différentes. Nous donnons en annexe A quelques statistiques (issues des données officielles de TREC) sur le contenu de cette collection (nombre de documents, de termes...). Le jeu de 50 requêtes sur lequel nous nous appuyons est celui utilisé lors de la campagne TREC-3 de 1994. Comme nous l'avons évoqué au chapitre 1 (*cf.* section 1.4.1), les requêtes sont représentées sous la forme de *topics* (dont un exemple est présenté en annexe A) composés de plusieurs champs. Dans notre cas, seuls les champs *titre* et *description* sont retenus. La longueur de ces requêtes correspond, pour la plupart d'entre elles, à deux ou trois phrases. Les fichiers de pertinence (qui associent à chaque requête la liste des documents qui ont été jugés au préalable comme pertinents ou non pertinents et qui est obtenue à l'aide d'une technique de *pooling* décrite en section 1.4.1) utilisés sont ceux de la campagne TREC.

3.4.2 Impact respectif des diverses informations linguistiques sur les performances des SRI

Pour évaluer l'apport individuel des 12 informations, nous nous appuyons sur l'architecture de test présentée précédemment. Nous intégrons en parallèle chacune des représentations linguistiques des documents et requêtes au sein du SRI. Ce dernier procède alors à l'appariement des index (un index correspond à une représentation particulière), évalue la pertinence de chaque document de la collection en fonction de la requête considérée et produit (pour chaque index) une liste de résultats qui correspond à l'ensemble des documents qu'il a retrouvé et classé par ordre décroissant de pertinence. À partir des 12 listes de résultats produites pour chaque requête, nous pouvons estimer les performances du SRI pour chaque type de représentation considéré en utilisant les

mesures d'évaluation communément utilisées en RI. Les résultats présentés en figure 3.3 indiquent la précision moyenne non interpolée (MAP) (définie en section 1.4.2) obtenue par le SRI pour chacun des index exploités (il s'agit plus précisément de la MAP moyennée sur les 50 requêtes de notre jeu de test).

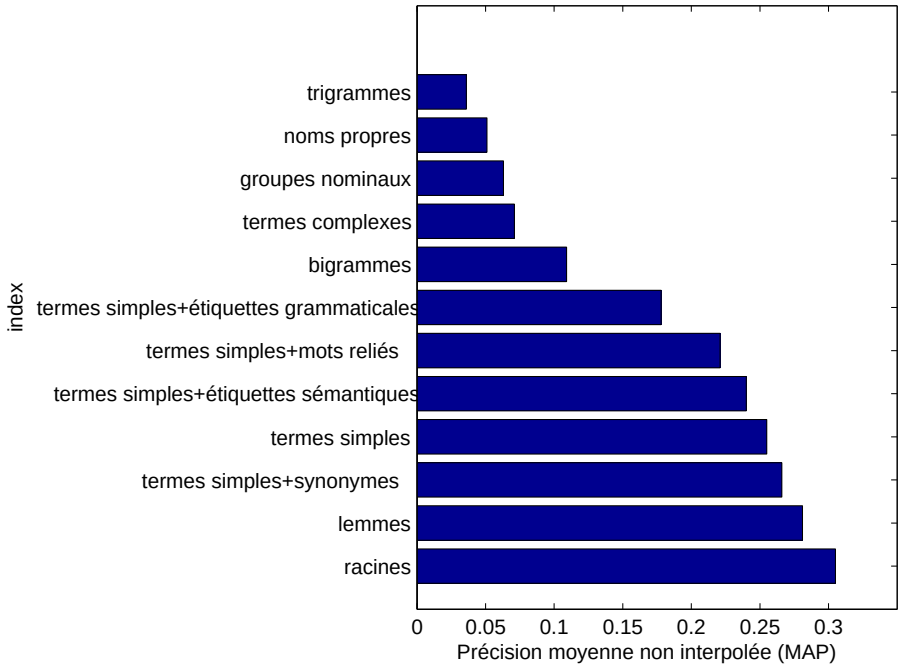

FIG. 3.3 – Performance du SRI pour chaque information linguistique manipulée

Les performances d'un SRI « standard », n'exploitant aucune information linguistique, correspondent sur la figure aux résultats obtenus par l'index des termes simples. Elles nous servent de référence pour évaluer l'apport des index « linguistiques ». L'analyse des résultats montre tout d'abord que les informations linguistiques prises en compte séparément ne contribuent pas toutes à l'amélioration des performances du système. En effet, seuls trois types de connaissances (les lemmes, les racines et les termes simples + leurs synonymes) permettent une augmentation de la MAP par rapport à celle d'un SRI traditionnel.

Il ressort tout d'abord de ces résultats que la prise en compte d'informations morphologiques s'avère intéressante en RI. Les outils utilisés pour leur acquisition étant particulièrement basiques, on peut supposer, comme le montreront les expériences présentées au chapitre 5, qu'en utilisant des méthodes plus évoluées (sans toutefois être plus complexes), les résultats peuvent être encore meilleurs. Concernant la comparaison de l'efficacité des lemmes par rapport aux pseudo-racines, les résultats obtenus donnent l'avantage aux méthodes de *stemming* pour notre corpus en anglais. Après une analyse

manuelle des documents retournés par ces deux types d'index, il apparaît que cette supériorité est principalement liée au fait que ces techniques permettent d'apparier des termes de catégories grammaticales différentes.

L'exploitation des informations de synonymie a également un impact positif. Là encore, ces connaissances sont obtenues avec des méthodes qui ne sont que partiellement efficaces. Le traitement de désambiguïsation des termes qu'elles imposent introduit du bruit (provoqué par l'assignation d'un sens incorrect aux termes), laissant penser qu'en recourant à des informations de synonymie plus fines, les performances pourraient être améliorées. De plus, en examinant les documents retournés par le SRI exploitant cet index, nous avons constaté que les synonymes associés aux termes simples des documents auraient certainement encore plus d'impact si leur forme était normalisée, c'est-à-dire si on procédait par exemple à leur racinisation.

Les autres informations de nature sémantique (mots reliés et étiquettes sémantiques) ne permettent pas au SRI d'obtenir des performances supérieures à celles d'un système classique à base de termes simples. Plusieurs raisons peuvent justifier ces résultats. Pour l'étiquetage sémantique (à chaque terme est associé un numéro correspondant à son sens dans WORDNET), la granularité trop fine des sens proposée dans WORDNET (*e.g.* les 63 sens associés au terme *break*) empêche l'appariement de deux termes pourtant sémantiquement liés mais possédant des étiquettes différentes (correspondant à des nuances de sens très fines). Pour les index contenant des termes simples associés à un ensemble de mots reliés, il semble que l'absence d'efficacité soit liée au fait que ces index sont souvent incomplets. En effet, au sein de WORDNET, les informations de morphologie dérivationnelle ne sont pas disponibles pour tous les termes présents dans la ressource. Seule une partie des termes des documents et requêtes sont par conséquent enrichis à l'aide de ces informations. De plus, les mots reliés ajoutés aux termes sont parfois non pertinents (*i.e.* non liés sémantiquement) ; ceci est dû principalement aux erreurs engendrées par le traitement de désambiguïsation des termes. D'une manière générale, l'un des principaux problèmes des informations issues de WORDNET est que les termes absents de la ressource sont laissés de côté (*i.e.* ramenés à leur forme lemmatisée ou à leur forme de base sans aucun autre traitement). L'enrichissement des représentations des documents et requêtes à l'aide de ces informations n'est donc que partiel.

La prise en compte d'informations morpho-syntaxiques (*i.e.* termes simples associés à leurs étiquettes grammaticales) donne des résultats décevants, certainement liés d'une part aux erreurs d'étiquetage de l'outil utilisé, mais également à la difficulté d'associer des étiquettes aux termes présents dans les titres des requêtes. Les titres sont en effet généralement représentés par un syntagme dépourvu notamment de ponctuation et de verbes.

Enfin, les résultats les plus décevants sont ceux obtenus en intégrant au sein du SRI des informations de nature syntaxique (bigrammes, termes complexes, groupes nominaux et trigrammes) puisque l'on constate une baisse très significative des performances des systèmes comparées à celles d'un SRI standard. La faiblesse de ces résultats ne semble pas liée à la mauvaise qualité des informations prises en compte. Une analyse manuelle des différentes représentations nous a en effet permis de constater que ces connaissances étaient plutôt pertinentes pour désigner les contenus textuels. Parmi les

différentes explications possibles, nous privilégions plutôt l'idée — au moins pour les termes complexes et les groupes nominaux — qu'en considérant les documents et requêtes uniquement à partir des informations syntaxiques qu'ils contiennent, on entraîne une sous-représentation de leur contenu textuel. En effet, les structures complexes, bien que réellement significatives, ne sont pas toujours très nombreuses au sein des documents et requêtes et, en particulier, dans une collection généraliste comme celle que nous utilisons pour ces expériences. Il apparaît donc primordial de les coupler avec d'autres connaissances (les termes simples par exemple). De plus, ces structures sont sujettes, comme nous l'avons vu au chapitre 2, à de multiples variations (typographiques, morpho-syntaxiques, syntaxiques...). Or, mis à part pour les termes complexes où l'outil employé (FASTR) détecte les variantes et les normalise, les autres méthodes utilisées pour extraire ces informations syntaxiques ne les prennent pas en compte.

D'une manière générale, le fait que certaines informations linguistiques, et plus particulièrement les informations d'ordre (morpho-)syntaxique et sémantique, ne fournissent pas de meilleurs résultats n'est pas surprenant. Nous avons volontairement adopté une représentation simpliste des documents pour pouvoir évaluer l'apport intrinsèque de chaque type de représentation et cela pénalise évidemment certaines représentations très spécialisées comme les trigrammes ou les noms propres. Ce n'est pas non plus problématique dans l'optique de leur couplage à condition que chaque index, quelles que soient ses performances, retrouve de façon complémentaire des documents pertinents. Dans le cas contraire où ces informations ne permettent pas au SRI de récupérer des documents différents de ceux retrouvés par un index plus performant, il apparaîtrait inutile de les exploiter pour le couplage puisqu'elles ne présenteraient aucune valeur ajoutée par rapport aux autres connaissances prises en compte.

Toutes ces observations mettent en évidence l'idée qu'on ne peut combiner des informations aussi diverses sans avoir une vue précise de la manière dont elles se comportent les unes par rapport aux autres. Seule une étude approfondie de leurs relations peut permettre en effet de savoir si les diverses informations exploitées sont complémentaires ou redondantes, si les différents gains observés sont susceptibles de s'additionner... Pour nous donner les moyens de mener une telle étude, nous avons mis en place une méthodologie, présentée ci-dessous, qui nous fournit dans notre cas des éléments de réponse quant à l'intérêt de coupler des informations linguistiques multi-niveaux.

3.4.3 Analyse des relations entre informations linguistiques multi-niveaux

Les deux expériences présentées ici cherchent à mettre en évidence les éventuelles relations susceptibles d'exister entre les informations linguistiques multi-niveaux intégrées au sein du SRI. La première vise plus particulièrement à analyser les corrélations entre les différentes listes de résultats retournées par le SRI. La seconde propose d'affiner cette analyse en étudiant les corrélations uniquement à partir des documents pertinents retrouvés par les systèmes.

3.4.3.1 Analyse des corrélations entre listes de résultats

Cette expérience propose d'étudier s'il existe des liens entre les résultats obtenus par le SRI qui intègre en parallèle les 12 représentations linguistiques des documents et requêtes. Pour cela, nous utilisons, pour chacun des 12 index, la liste correspondante des documents retournés par le système en réponse aux requêtes. Pour faciliter l'analyse des relations entre ces listes, nous les comparons deux à deux, en nous appuyant sur les rangs assignés à chaque document dans les listes ordonnées de résultats.

Nous mesurons donc, pour chaque requête, la corrélation entre deux listes de documents (notées l_1 et l_2), chaque liste correspondant aux résultats obtenus par le SRI qui intègre une représentation linguistique particulière des textes et questions. Nous cherchons à évaluer si ces listes sont similaires (*i.e.* les documents retrouvés dans un ordre identique ou proche) ou non. Pour ce faire, nous utilisons le cœfficient de corrélation de rang de Spearman (noté ρ), qui examine s'il existe une relation entre les rangs des documents retrouvés par les deux listes, et qui est défini par la formule suivante :

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

où :
D est la distance entre les rangs d'un document dans l_1 et l_2,
N est le nombre de paires de valeurs analysées (*i.e.* le nombre de documents).

Cette mesure étant calculée pour une requête donnée, nous répétons l'opération pour les 50 requêtes de notre collection et obtenons la corrélation moyenne entre les listes de deux index. Ces calculs sont effectués pour toutes les paires d'index. Plus le cœfficient est proche de +1, plus les listes de documents sont corrélées (*i.e.* similaires).

Pour pouvoir être appliquée, cette mesure exige que l_1 et l_2 contiennent exactement les mêmes documents (seul leur rang peut différer). Or dans notre cas, il apparaît fréquemment qu'un index retrouve par le biais du SRI des documents non ramenés par l'autre index (et réciproquement). Ce type de problème a déjà été étudié en RI et plusieurs propositions d'améliorations ont été faites (Fagin *et al.*, 2003; Bar-Ilan *et al.*, 2004). Pour notre part, nous le contournons en modifiant l_1 et l_2 afin de garder uniquement les documents présents dans l'intersection des deux listes. Si cette intersection est trop petite, calculer le cœfficient de corrélation n'aurait guère de sens. Ainsi, dans les expériences présentées ci-après, nous ne calculons la corrélation que si l'intersection des deux listes de résultats considérées compte plus de 200 documents. Si ce n'est pas le cas, cela signifie que les index renvoient des résultats trop différents et on peut considérer qu'ils ne sont pas corrélés.

La figure 3.4 présente la moyenne des cœfficients obtenus sur les 50 requêtes pour les paires d'index dont l'intersection des listes de résultats contient en moyenne plus de 200 documents. La corrélation est indiquée en abcisse et la taille moyenne de l'intersection est donnée sur la ligne. Pour une raison de lisibilité, *synonymes* signifie « termes simples + un ensemble de synonymes » et *mots reliés* correspond aux termes simples associés à un ensemble de mots unis par un lien morpho-sémantique.

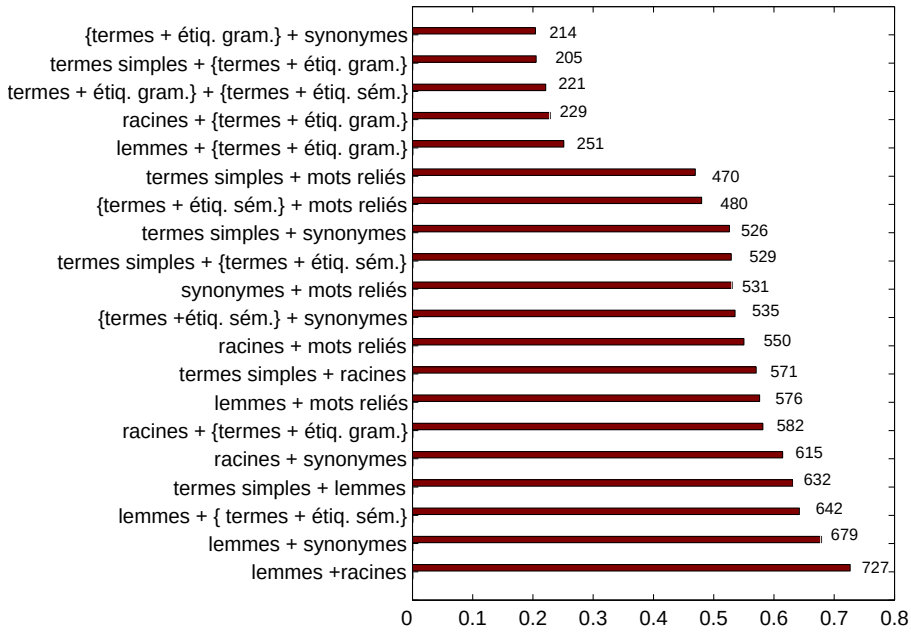

FIG. 3.4 – Moyenne des corrélations par couple d'index (sur 50 requêtes)

Ces résultats conduisent à plusieurs remarques. Une première observation, peu surprenante, est qu'aucune paire d'index n'est parfaitement corrélée. Seules vingt paires proposent des listes de résultats comptant plus de 200 documents en commun, et parmi celles-ci, les cœfficients sont tous inférieurs à 0.8. Parmi les paires d'index qui ont les cœfficients les plus élevés (*i.e.* compris entre 0.6 et 0.8), les informations linguistiques les plus liées sont, sans surprise, des index couplant des informations qui appartiennent au même niveau de langue et plus précisément au niveau morphologique, *e.g.* les paires *lemmes-racines, lemmes-termes simples, racines-termes simples*... Le lien unissant des informations d'ordre sémantique semble moins évident puisque la première paire la plus corrélée (le couple *synonymes-mots reliés*) a un cœfficient inférieur à 0.6. Ce sont plutôt les paires combinant des connaissances morphologique et sémantique qui apparaissent comme les plus unies (*e.g.* les associations *lemmes-synonymes, racines-synonymes* ou *lemmes-étiquettes sémantiques*), ce qui est assez attendu puisque nous nous appuyons sur les termes simples (ou les lemmes) pour extraire les informations sémantiques de WORDNET. Ces premiers résultats semblent donc indiquer que le couplage d'informations appartenant aux niveaux de langue morphologiques et sémantiques n'est pas nécessairement pertinent puisque l'on risque d'obtenir des résultats assez redondants.

Parmi les index qui ne sont corrélés à aucun autre (*i.e.* qui produisent, par le biais du SRI qui les intègre, des listes de documents très différentes) apparaissent en majorité ceux manipulant une information syntaxique (bigrammes, trigrammes...). En nous appuyant sur les résultats individuels de ces informations (*cf.* la MAP présentée en section précédente), nous pouvons néanmoins expliquer cette différence par le fait que ces

informations linguistiques ne retrouvent, contrairement aux autres index, que très peu de documents pertinents, ce qui mène à des listes de résultats très dissemblables.

Les corrélations présentées ici nous informent uniquement de la similarité des listes de résultats retournés par le SRI intégrant deux types d'index. Elles n'intègrent cependant aucune information sur la pertinence des documents présents au sein de ces listes. Or, il est important dans notre cas de savoir si la relation qui unit deux index correspond au fait qu'ils permettent au SRI de retrouver des documents pertinents ou non. Nous souhaitons également déterminer de manière plus précise la nature de la relation qui unit deux informations linguistiques, *i.e.* s'il s'agit d'une relation de redondance ou de complémentarité. Nous proposons donc dans la section suivante de nouvelles expérimentations qui visent à analyser les corrélations entre les informations linguistiques du point de vue de leur capacité à retrouver ou non des documents pertinents.

3.4.3.2 Analyse des corrélations entre listes de documents pertinents

Cette nouvelle série d'expérimentations se déroule en plusieurs étapes. La première expérience vise à analyser les relations entre les différents index uniquement à partir de leur capacité à retrouver (ou non) les documents pertinents. En s'appuyant sur la liste des documents qui devraient idéalement être ramenés par le SRI pour une requête donnée[8], il s'agit d'examiner si deux index ont un comportement similaire ou différent pour retrouver ou ne pas retrouver les documents de cette liste. À la suite des résultats obtenus, deux points nécessitent d'être approfondis par le biais d'expériences complémentaires. Le premier concerne plus précisément les cas où les index semblent se comporter de manière similaire pour retrouver (ou non) les mêmes documents pertinents (*i.e.* les informations linguistiques semblent redondantes). Nous souhaitons déterminer si les informations linguistiques sont liées parce qu'elles permettent de retourner toutes deux les documents pertinents ou au contraire parce qu'elles ne les retrouvent pas. Le second point s'intéresse aux cas où les index ont un comportement différent dans la récupération des documents pertinents (ils sont considérés alors comme complémentaires). Il s'agit de vérifier la nature exacte de cette complémentarité, *i.e.* savoir si la différence observée entre deux index correspond véritablement au fait qu'ils retrouvent tous deux des documents pertinents différents ou si elle correspond au contraire au fait que seul un des deux index est performant pour leur récupération.

Nous revenons donc successivement sur ces trois expériences. Nous présentons, pour chacune d'elles, les différents résultats obtenus sur notre collection de test, résultats sur lesquels nous nous appuyons pour tirer un certain nombre de conclusions.

Analyse des similarités entre index pour la récupération de documents pertinents

Cette expérience s'appuie, pour analyser les relations entre les diverses informations linguistiques prises en compte, uniquement sur leur capacité à retrouver ou

[8] Cette liste est obtenue à partir des informations présentes dans les fichiers de pertinence de TREC.

non, par le biais du SRI, les documents pertinents de la collection pour une requête donnée. Pour cela, nous proposons d'utiliser une méthode binaire simple qui consiste à évaluer si, à partir de la liste des documents pertinents pour chaque requête fournie avec la collection TIPSTER, les index (évalués par paires) permettent au SRI de retrouver ou non ces documents pertinents. Pour une requête donnée, un document est jugé retrouvé par un index s'il est proposé dans les 1000 premières réponses.

À partir de la liste des documents jugés *a priori* pertinents pour une requête donnée et pour toutes les paires d'informations linguistiques prises en compte, nous notons 1 si le premier index (nommé $index_1$) de la paire (*resp.* le second index de la paire noté $index_2$) retrouve le document pertinent et 0 sinon. Nous procédons de cette façon pour l'ensemble des documents de la liste, des requêtes et des paires d'index. La figure 3.5 illustre le type de résultat obtenu pour une paire d'index donnée.

requête	liste des documents pertinents issue des fichiers de pertinence	SRI intégrant l'index des lemmes	SRI intégrant l'index des bigrammes
151	WSJ870402-0112	1	0
151	WSJ870402-0118	1	0
151	WSJ870413-0098	0	0
151	WSJ870414-0098	1	0
151	WSJ870420-0003	0	0
151	WSJ870421-0019	0	0
151	WSJ870422-0137	0	0
151	WSJ870428-0145	1	1
151	WSJ870428-0147	0	0
151	WSJ870429-0078	0	0
151	WSJ870430-0121	0	0

FIG. 3.5 – Exemple de liste de documents pertinents retrouvés (1) ou non (0) par deux index pour une requête donnée

Nous cherchons ensuite à évaluer, pour chaque paire d'informations linguistiques étudiée, la similarité entre les résultats obtenus par $index_1$ et $index_2$. Autrement dit, nous souhaitons savoir si les documents pertinents retrouvés par le premier index sont les mêmes que ceux retournés par le second, et réciproquement. Pour juger de la ressemblance de ces deux listes, nous utilisons l'indice de similarité suivant :

$$sim = \frac{n - \sum_{i=1}^{n} (index1_i - index2_i)^2}{n} \times 100$$

où[9] :
n est le nombre de paires de valeurs analysées,
$index1_i$ et $index2_i$ représentent les valeurs (dans notre cas binaires) retournées par

[9]Cet indice de similarité est obtenu en soustrayant à n l'indice de dissimilarité représenté par la distance euclidienne au carré : $\sum_{i=1}^{n} (index1_i - index2_i)^2$. La division de $n - \sum_{i=1}^{n} (index1_i - index2_i)^2$ par n nous permet d'obtenir des valeurs normalisées.

l'$index_1$ et l'$index_2$ pour un document pertinent i.

Cet indice de similarité permet de mesurer le taux de réponses identiques entre les deux index. Un taux proche de 100% signifie que les deux index d'une paire retrouvent (ou ne retrouvent pas), par le biais du SRI, les mêmes documents pertinents. Leurs listes de résultats respectives sont donc similaires. Un taux proche de 0% implique que les deux index sont complémentaires puisque les documents pertinents retrouvés par l'un ne sont pas retournés par l'autre (et inversement). La figure 3.6 représente la moyenne (pour les 50 requêtes) des taux de réponses identiques obtenues pour chaque paire d'index. Les résultats sont classés par ordre décroissant, *i.e.* des paires possédant des listes de résultats très similaires à celles ayant des listes très différentes.

L'analyse des différents taux obtenus montre que les paires d'index qui produisent des résultats très similaires (dont le taux de similarité est $>$ à 90%) sont celles qui combinent généralement des informations de même niveau de langue et, plus particulièrement, d'ordre syntaxique (*e.g.* les paires *groupes nominaux-trigrammes, bigrammes-termes complexes, groupes nominaux-termes complexes*) et celles couplant des informations morphologique et sémantique (*e.g.* les couples *lemmes-synonymes, racines-synonymes, lemmes-étiquettes sémantiques*). Pour les paires associant uniquement des informations morphologiques, le couple *lemmes-racines* retourne également des résultats assez proches (identiques pour 91,17%). Pour la paire *racines-termes simples*, la différence est légèrement plus importante puisque dans plus de 15% des cas ces deux informations fournissent des résultats complémentaires. Les couples composés uniquement de connaissances sémantiques (*e.g.* les paires *synonymes-étiquettes sémantiques, synonymes-mots reliés, étiquettes sémantiques-mots reliés*) semblent également avoir des comportements assez proches pour retrouver les documents pertinents (leur taux de similarité avoisine les 85%). Les informations d'ordre à la fois morphologique et syntaxique, *e.g.* les paires *termes simples-termes complexes, lemmes-bigrammes, racines-termes complexes...*, semblent assez complémentaires puisque leur taux de similarité ne dépasse jamais les 75%, et est le plus souvent proche des 60%. Enfin, parmi les taux de similarité les plus bas (*i.e.* $<$ à 40%), on trouve essentiellement des couples qui combinent des informations syntaxiques et sémantiques, *e.g.* les paires *trigrammes-synonymes, trigrammes-étiquettes sémantiques, trigrammes-mots reliés, groupes nominaux-synonymes, groupes nominaux-étiquettes sémantiques...* Ces deux types de connaissances semblent donc avoir un comportement complémentaire pour retrouver (ou non) les documents pertinents.

L'expérience suivante propose de compléter ces observations en cherchant à déterminer si les couples d'index qui semblent redondants le sont parce qu'ils retrouvent tous deux des documents pertinents ou, au contraire, parce qu'ils échouent à les retrouver.

Analyse des cas de redondances

Pour les cas où les résultats renvoyés par deux index sont similaires ou presque, nous proposons de mesurer le taux de documents pertinents effectivement retrouvés par les deux index, et le taux de ceux non retrouvés. Les résultats obtenus sont présentés en

paire d'index	taux de résultats identiques (en %)
groupes nominaux + trigrammes	95,86
bigrammes + termes complexes	93,96
groupes nominaux + termes complexes	93,15
lemmes + synonymes	91,79
lemmes + racines	91,17
termes complexes + trigrammes	90,96
bigrammes + groupes nominaux	89,73
racines + synonymes	89,31
bigrammes + trigrammes	88,23
noms propres + trigrammes	87,92
lemmes + étiquettes sémantiques	87,49
lemmes + termes simples	87,07
étiquettes sémantiques + synonymes	86,68
noms propres + groupes nominaux	86,43
synonymes + mots reliés	85,43
racines + étiquettes sémantiques	85,36
termes simples + synonymes	85,03
lemmes + mots reliés	84,71
racines + termes simples	84,49
termes simples + étiquettes sémantiques	83,67
racines + mots reliés	83,48
noms propres + termes complexes	82,74
étiquettes sémantiques + mots reliés	82,42
termes simples + mots reliés	81,98
noms propres + bigrammes	81,39
termes complexes + étiquettes grammaticales	79,51
bigrammes + étiquettes grammaticales	79,13
étiquettes grammaticales + mots reliés	78,33
groupes nominaux + étiquettes grammaticales	78,02
lemmes + étiquettes grammaticales	77,22
étiquettes grammaticales + trigrammes	77,02
noms propres + étiquettes grammaticales	76,68
étiquettes grammaticales + étiquettes sémantiques	76,62
termes simples + étiquettes grammaticales	75,54
termes simples + termes complexes	75,54
racines + étiquettes grammaticales	75,43
étiquettes grammaticales + synonymes	75,36
termes simples + bigrammes	65,77
noms propres + termes simples	64,92
lemmes + bigrammes	64,45
racines + bigrammes	63,38
lemmes + termes complexes	62,99
noms propres + lemmes	62,64
racines + termes complexes	61,91
termes simples +groupes nominaux	61,77
noms propres + racines	61,56
termes simples + trigrammes	59,77
lemmes + groupes nominaux	59,40
bigrammes + étiquettes sémantiques	58,91
racines + groupes nominaux	58,46
bigrammes + mots reliés	58,04
lemmes + trigrammes	57,34
racines + trigrammes	56,69
bigrammes + synonymes	55,15
termes complexes + mots reliés	50,46
termes complexes + étiquettes sémantiques	49,49
termes complexes + synonymes	45,81
noms propres + mots reliés	44,02
groupes nominaux + mots reliés	43,69
noms propres + étiquettes sémantiques	43,62
groupes nominaux + étiquettes sémantiques	42,07
noms propres + synonymes	40,16
groupes nominaux + synonymes	38,03
trigrammes + mots reliés	33,69
trigrammes + étiquettes sémantiques	33,27
trigrammes + synonymes	27,77

FIG. 3.6 – Similitude (en %) des résultats (documents pertinents retrouvés (ou non)) pour deux index d'une paire

figure 3.7. Ce tableau se lit de la manière suivante : pour le premier couple d'index *groupes nominaux-trigrammes* par exemple, les listes de résultats retournées par ces deux index sont similaires à 95,86% (colonne 2). Lorsque ces deux index renvoient les mêmes résultats, pour 99% des cas (colonne 4), les documents pertinents n'ont pas été retrouvés, et inversement, pour 1% des cas (colonne 3), les documents pertinents ont effectivement été retournés.

À partir de ces données, nous distinguons principalement deux groupes : les paires qui combinent uniquement des informations syntaxiques (*e.g.* les couples *groupes nominaux-trigrammes, bigrammes-termes complexes, groupes nominaux-termes complexes*...) et les paires qui manipulent des informations morpho-sémantiques (voire seulement sémantiques), telles que les couples *lemme-synonymes, racine-synonymes, lemmes-étiquettes sémantiques, synonymes-mots reliés*. Pour le premier groupe, il apparaît nettement que, pour notre collection, les informations syntaxiques, lorsqu'elles sont couplées, ne permettent au SRI de retrouver que très peu de documents pertinents (le couple le plus performant est la paire *bigrammes-termes complexes* qui, sur les 93% de résultats identiques que ces deux index fournissent, ramène seulement 25% de documents pertinents). D'une manière générale, nous observons que le couplage d'informations de nature essentiellement syntaxique apparaît peu pertinent puisque ces dernières se comportent souvent de façon similaire et que cette similarité correspond la plupart du temps au fait qu'elles ne retrouvent que très rarement des documents pertinents. Inversement, et contrairement à ce que pouvaient laisser penser les premières analyses de corrélation sur les listes de documents initialement retrouvés par les index (section 3.4.3.1), les résultats obtenus par la combinaison de connaissances sémantiques ou morpho-sémantiques se révèlent plus positifs puisqu'ils correspondent en grande partie à la récupération de documents pertinents.

Nous venons d'analyser les cas de redondance des index. Dans le but de proposer une analyse des relations qui soit aussi complète que possible, nous nous intéressons à présent plus en détail aux cas de complémentarité.

Analyse des cas de complémentarité entre index

Nous examinons maintenant les cas où deux index ont un pourcentage élevé de réponses différentes lorsque l'on analyse leur capacité à retrouver les documents pertinents. Nous cherchons plus précisément à déterminer si ces index se complètent et retrouvent tous deux des documents pertinents différents, ou si ce pourcentage est simplement lié au fait qu'un seul des deux index est efficace pour retrouver les bons documents.

Pour cela, nous proposons de mesurer un indice de complémentarité entre les deux index, obtenu en calculant, pour les cas où les listes de résultats sont différentes, le taux de documents pertinents retrouvés par le premier index et non retrouvés par le second et le taux de documents pertinents ramenés par le second index et non par le premier. Les résultats obtenus sont répertoriés dans le tableau de la figure 3.8. Ils représentent la contribution de chacun des index d'une paire pour retrouver les documents pertinents.

paire d'index	taux de résultats identiques (en %)	taux de documents pertinents retrouvés par les 2 index (en %)	taux de documents pertinents non retrouvés par les 2 index (en %)
groupes nominaux + trigrammes	95,86	1	99
bigrammes + termes complexes	93,96	25	75
groupes nominaux + termes complexes	93,15	12	88
lemmes + synonymes	91,79	79	21
lemmes + racines	91,17	80	20
termes complexes + trigrammes	90,96	1	99
bigrammes + groupes nominaux	89,73	17	83
racines + synonymes	89,31	80	20
bigrammes + trigrammes	88,23	1	99
noms propres + trigrammes	87,92	1	99
lemmes + étiquettes sémantiques	87,49	77	23
lemmes + termes simples	87,07	77	23
étiquettes sémantiques + synonymes	86,68	76	24
noms propres + groupes nominaux	86,43	5	95
synonymes + mots reliés	85,43	75	25
racines + étiquettes sémantiques	85,36	78	22
termes simples + synonymes	85,03	77	23
lemmes + mots reliés	84,71	77	23
racines + termes simples	84,49	78	22
termes simples + étiquettes sémantiques	83,67	75	25
racines + mots reliés	83,48	78	22
noms propres + termes complexes	82,74	17	83
étiquettes sémantiques + mots reliés	82,42	73	27
termes simples + mots reliés	81,98	74	26
noms propres + bigrammes	81,39	25	85
termes complexes + étiquettes grammaticales	79,51	30	70
bigrammes + étiquettes grammaticales	79,13	44	56
étiquettes grammaticales + mots reliés	78,33	64	36
groupes nominaux + étiquettes grammaticales	78,02	21	79
lemmes + étiquettes grammaticales	77,22	72	28
étiquettes grammaticales + trigrammes	77,02	5	95
noms propres + étiquettes grammaticales	76,68	22	88
étiquettes grammaticales + étiquettes sémantiques	76,62	66	34
termes simples + étiquettes grammaticales	75,54	68	32
termes simples + termes complexes	75,54	44	56
racines + étiquettes grammaticales	75,43	73	27
étiquettes grammaticales + synonymes	75,36	70	30
termes simples + bigrammes	65,77	57	43
noms propres + termes simples	64,92	41	59
lemmes + bigrammes	64,45	64	36
racines + bigrammes	63,38	65	35
lemmes + termes complexes	62,99	49	51
noms propres + lemmes	62,64	42	58
racines + termes complexes	61,91	51	49
termes simples +groupes nominaux	61,77	31	69
noms propres + racines	61,56	44	56
termes simples + trigrammes	59,77	10	90
lemmes + groupes nominaux	59,40	38	62
bigrammes + étiquettes sémantiques	58,91	56	44
racines + groupes nominaux	58,46	38	62
bigrammes + mots reliés	58,04	54	46
lemmes + trigrammes	57,34	13	87
racines + trigrammes	56,69	14	86
bigrammes + synonymes	55,15	61	39
termes complexes + mots reliés	50,46	40	60
termes complexes + étiquettes sémantiques	49,49	42	58
termes complexes + synonymes	45,81	47	53
noms propres + mots reliés	44,02	36	64
groupes nominaux + mots reliés	43,69	28	72
noms propres + étiquettes sémantiques	43,62	39	61
groupes nominaux + étiquettes sémantiques	42,07	30	70
noms propres + synonymes	40,16	44	56
groupes nominaux + synonymes	38,03	35	65
trigrammes + mots reliés	33,69	8	92
trigrammes + étiquettes sémantiques	33,27	10	90
trigrammes + synonymes	27,77	12	88

FIG. 3.7 – Taux de documents pertinents identiques retrouvés (colonne 3) ou non retrouvés (colonne 4) simultanément par les deux index

paire d'index (index 1 + index 2)	taux de résultats différents (en %)	documents pertinents retrouvés uniquement par le 1er index (en %)	documents pertinents retrouvés uniquement par le 2nd index (en %)
trigrammes + synonymes	72,22	1	99
trigrammes + étiquettes sémantiques	66,72	1	99
trigrammes + mots reliés	66,30	2	98
groupes nominaux + synonymes	61,96	0	100
noms propres + synonymes	59,83	4	96
groupes nominaux + étiquettes sémantiques	57,92	2	98
noms propres + étiquettes sémantiques	56,37	6	94
groupes nominaux + mots reliés	56,30	2	98
noms propres + mots reliés	55,97	8	92
termes complexes + synonymes	54,18	2	98
termes complexes + étiquettes sémantiques	50,50	3	97
termes complexes + mots reliés	49,53	4	96
bigrammes + synonymes	44,86	8	92
racines + trigrammes	43,31	97	3
lemmes + trigrammes	42,66	98	2
bigrammes + mots reliés	41,95	13	87
racines + groupes nominaux	41,54	99	1
bigrammes + étiquettes sémantiques	41,08	10	90
lemmes + groupes nominaux	40,60	99	1
termes simples + trigrammes	40,23	97	3
noms propres + racines	38,43	8	92
termes simples +groupes nominaux	38,22	97	3
racines + termes complexes	38,09	95	5
noms propres + lemmes	37,35	8	92
lemmes + termes complexes	37,01	96	4
racines + bigrammes	36,62	92	8
lemmes + bigrammes	35,55	93	7
noms propres + termes simples	35,07	9	91
termes simples + bigrammes	34,23	91	9
étiquettes grammaticales + synonymes	24,63	9	91
racines + étiquettes grammaticales	24,57	92	8
termes simples + étiquettes grammaticales	24,46	86	14
termes simples + termes complexes	24,46	86	14
étiquettes grammaticales + étiquettes sémantiques	23,37	18	82
noms propres + étiquettes grammaticales	23,31	25	75
étiquettes grammaticales + trigrammes	22,98	96	4
lemmes + étiquettes grammaticales	22,78	94	6
groupes nominaux + étiquettes grammaticales	21,98	95	5
étiquettes grammaticales + mots reliés	21,66	20	80
bigrammes + étiquettes grammaticales	20,87	25	75
termes complexes + étiquettes grammaticales	20,49	15	85
noms propres + bigrammes	18,60	47	53
termes simples + mots reliés	18,01	61	39
étiquettes sémantiques + mots reliés	17,57	56	44
noms propres + termes complexes	17,25	58	42
racines + mots reliés	16,51	84	16
termes simples + étiquettes sémantiques	16,32	56	44
racines + termes simples	15,51	64	36
lemmes + mots reliés	15,28	83	17
termes simples + synonymes	14,96	38	62
racines + étiquettes sémantiques	14,63	81	19
synonymes + mots reliés	14,56	75	25
noms propres + groupes nominaux	13,56	80	20
étiquettes sémantiques + synonymes	13,31	30	70
lemmes + termes simples	12,93	60	40
lemmes + étiquettes sémantiques	12,50	82	18
bigrammes + trigrammes	11,77	100	0
racines + synonymes	10,68	68	32
bigrammes + groupes nominaux	10,27	95	5
termes complexes + trigrammes	9,04	93	7
lemmes + racines	8,83	46	54
lemmes + synonymes	8,20	67	33
groupes nominaux + termes complexes	6,84	89	11
bigrammes + termes complexes	6,04	83	17
groupes nominaux + trigrammes	4,13	20	80

FIG. 3.8 – Pourcentage de documents pertinents retrouvés par le 1er index de la paire et non retrouvés par le 2nd (colonne 3) et pourcentage de documents pertinents retrouvés par le 2nd index de la paire et non retrouvés par le 1er (colonne 4)

Deux index sont donc considérés comme complémentaires si les taux présents dans les colonnes 3 et 4 sont plus ou moins proches de 50%. Le fait qu'une des deux colonnes ait un taux élevé et l'autre un taux faible (*e.g.* le couple *trigrammes-synonymes* de la première ligne) signifie que pour les cas où les listes de résultats fournies par ces deux index sont différentes (dans 72% (colonne 2) des cas pour le même exemple), seul l'un des deux index est performant pour retrouver les documents pertinents (pour ce couple, seul l'index des synonymes est efficace). Il existe bien une complémentarité entre ces deux index (l'un retrouve toujours les documents pertinents, l'autre non) mais elle ne répond pas à nos objectifs, puisque nous cherchons uniquement à repérer les informations linguistiques dont le couplage permet au SRI de retrouver davantage de documents pertinents.

D'après les résultats obtenus, il apparaît d'une manière générale que le nombre d'index qui se révèlent complémentaires pour retrouver des documents pertinents est assez faible. Les couples d'informations linguistiques qui ont un taux proche de 50% concernent les index dont les listes de documents sont plus fréquemment identiques que différentes (les taux de résultats différents ne dépassent pas les 18%). Nous pouvons néanmoins faire un certain nombre d'observations intéressantes, suite à cette étude, sur ces cas de complémentarité. Afin d'être le plus synthétique possible, nous découpons notre analyse en étudiant d'une part l'intérêt de coupler des informations mono-niveau et, d'autre part, l'impact du couplage de connaissances multi-niveaux.

Pour les informations morphologiques, il apparaît tout d'abord que le couplage peut parfois permettre au SRI de retrouver davantage de documents pertinents. En effet, bien que ces informations semblent le plus souvent retourner des résultats identiques, on constate une véritable complémentarité de ces connaissances pour les cas où les documents pertinents retrouvés sont différents (pour moins de 15%). Ainsi, pour le couple *racines-termes simples*, dans 64% des cas les racines permettent de retrouver des documents pertinents non retournés par les termes simples et inversement. Ces résultats s'expliquent en partie par le fait que les outils utilisés pour le *stemming* provoquent des erreurs (*e.g.* erreurs de sur-racinisation ou sous-racinisation) compensées par la prise en compte des termes simples, et inversement. Ce constat est valable également pour la paire *lemmes-termes simples*. Le couplage *lemmes-racines* semble également pertinent : les racines permettent de retrouver dans 54% des cas des documents non retrouvés par les lemmes, et réciproquement. Un examen manuel des résultats montre que la prise en compte des lemmes est notamment efficace lorsque les formes des mots sont irrégulières (*e.g.* les formes *giv* et *gav* qui ne peuvent être appariées avec les outils de racinisation classique). Les racines sont quant à elles plus performantes pour mettre en correspondance des formes appartenant à des catégories grammaticales différentes (*e.g.* *retrieval* et *retrieve*).

Pour le couplage d'informations syntaxiques, ce sont davantage des cas de non-complémentarité qui se détachent de ces résultats. Il apparaît tout d'abord que la combinaison des bigrammes et trigrammes ne s'avère pas pertinente puisque, pour notre collection, les documents retrouvés en utilisant des représentations de trigrammes sont également retournés par l'index des bigrammes. Les trigrammes, bien que globalement inefficaces, peuvent dans quelques cas être utiles pour retrouver des documents pertinents

non retournés par les index de termes complexes ou de groupes nominaux. Ces derniers sont en effet parfois limités par leur méthode d'acquisition, qui ne permet pas toujours de détecter toutes les structures complexes présentes dans les documents. Les quelques cas de complémentarité existant entre les informations syntaxiques, notamment pour les paires *groupes nominaux-termes complexes*, *bigrammes-termes complexes*, sont également liés aux méthodes utilisées pour extraire des textes ces diverses informations. Les termes complexes, par exemple, sont obtenus à l'aide de patrons morpho-syntaxiques. La liste de ces patrons n'étant pas exhaustive (et l'étiquetage morpho-syntaxique pas toujours exempt d'erreurs), certains termes complexes ne sont par conséquent pas identifiés alors qu'ils le sont avec des outils basés sur des méthodes numériques.

Le couplage de connaissances sémantiques, et plus particulièrement celles issues des ressources pré-construites, peut également s'avérer intéressant. Les informations de synonymie et de liens morpho-sémantiques (mots reliés) issues de WORDNET sont ainsi souvent complémentaires. Les mots reliés permettent d'étendre les mots des textes en utilisant des informations sémantiques inter-catégorielles, à la différence des synonymes qui sont souvent plus nombreux pour enrichir les représentations mais limités à une seule catégorie grammaticale. La combinaison de ces deux connaissances permet donc une caractérisation plus riche des contenus textuels.

Ces différentes remarques concernaient le couplage d'informations appartenant à un même niveau de langue. Pour la combinaison de connaissances multi-niveaux, plusieurs remarques émergent des résultats obtenus. Les informations morphologiques et sémantiques apparaissent tout d'abord comme les connaissances les plus intéressantes à combiner du point de vue de leur complémentarité. Bien que les taux soient encore un peu faible[10], le recours conjoint à des index morphologiques et aux synonymes ou aux mots reliés permet de retrouver plus de documents que chaque index pris individuellement. Leur couplage peut constituer une piste d'autant plus intéressante à explorer que les résultats individuels de ces index sont plutôt bons par rapport aux traditionnels termes simples et peuvent certainement encore être améliorés. Ces observations mettent également en évidence l'intérêt des expérimentations menées ici, et plus généralement, d'étudier en profondeur les relations entre les informations linguistiques, puisque, comme nous l'avons déjà remarqué, les premières analyses de corrélations réalisées — à l'aide du cœfficient de Spearman — (proposées en section 3.4.3.1) n'avaient pas permis dans un premier temps de détecter ces cas de complémentarité.

L'usage conjoint d'informations syntaxiques et sémantiques (noms propres mis à part) ne présente que très peu de cas de complémentarité, puisque la plupart du temps, seules les connaissances sémantiques permettent au SRI de retrouver les documents pertinents (*e.g.* pour les paires *trigrammes-synonymes, groupes nominaux-étiquettes sémantiques, termes complexes-mots reliés*...). Les rares fois où la tendance est inversée s'expliquent par le fait que certains termes des documents sont absents des ressources sémantiques, mais apparaissent dans les index « syntaxiques » (*e.g.* pour la paire *bigrammes-étiquettes sémantiques*).

[10] Pour le couplage des index « simples » (*i.e.* exploitant soit des lemmes ou des racines) avec un index de synonymes, le taux de complémentarité avoisine les 30-35% pour un taux de résultats différents pour ces deux index d'environ 10%.

Les noms propres ont un comportement différent des autres informations sémantiques (dû à leur méthode d'acquisition essentiellement morpho-syntaxique). D'une manière générale, leur couplage avec des informations syntaxiques est efficace (*e.g.* par exemple le couple *noms propres-bigrammes*). Ceci s'explique en particulier par le fait que même si représenter des documents uniquement par des noms propres qu'ils contiennent conduit à une sous-représentation de leur contenu informationnel, ces noms propres capturent des termes simples (par exemple *Japan*) que ne peuvent détecter les index « syntaxiques ». L'intégration combinée des deux types de connaissances est donc intéressante. Concernant le couplage d'informations morphologiques et syntaxiques, les cas de complémentarité sont également peu nombreux. D'une manière générale, les connaissances d'ordre morphologique ont un impact nettement supérieur à celui des informations syntaxiques sur les performances des SRI ; il apparaît donc normal que la plupart des documents pertinents soient retrouvés par leur biais.

Ces diverses expériences ont cherché à évaluer les relations susceptibles d'exister entre des paires d'index utilisant des informations linguistiques différentes. Afin d'avoir un aperçu plus global de la corrélation entre l'ensemble des connaissances prises en compte, nous proposons de terminer cette série d'expérimentations en appliquant une méthode de classification sur les résultats retrouvés par les index. Notre objectif est d'évaluer si des classes d'informations linguistiques ayant un impact similaire sur les résultats des SRI peuvent être formées.

3.4.4 Classification des informations linguistiques selon leur impact en RI

Les analyses précédentes des corrélations manipulaient des index d'informations linguistiques par paires. Pour cette dernière expérience, nous souhaitons aller plus loin et étudier de manière simultanée les relations entre tous les index. Pour cela, nous proposons d'appliquer un algorithme de classification ascendante hiérarchique sur les listes de résultats produites par chaque type d'index afin d'essayer de former des classes d'informations linguistiques en fonction des documents pertinents (et de leur positions respectives dans la liste des résultats) qu'ils permettent au SRI de retrouver. Les données sur lesquelles s'applique cet algorithme sont construites à partir de la liste des documents qui doivent idéalement être retrouvés par le SRI pour une requête donnée. Pour chacun de ces documents pertinents, on regarde à quel rang chaque index pris individuellement a retrouvé ce document. On procède de la sorte pour tous les documents de la liste et pour tous les index. Lorsqu'un index n'a pas retrouvé le document pertinent, un rang fictif lui est attribué, qui correspond au rang maximal des documents retrouvés par le SRI pour cet index. On obtient à la suite de ce traitement une matrice de données, dont la figure 3.9 présente un exemple, sur laquelle est appliqué l'algorithme de classification ascendante hiérarchique, utilisé avec comme critères la distance euclidienne et l'agrégation selon la méthode de Ward[11]. Le dendrogramme obtenu est présenté en figure 3.10.

[11] Pour le paramétrage de l'algorithme, nous avons expérimenté plusieurs autres mesures et techniques d'agrégation, les résultats obtenus apparaissant toutefois très similaires à ceux présentés ici.

req.	document pertinent	lemmes	racines	termes simples	étiquettes gram-maticales	bi-grammes	groupes nominaux	termes complexes	tri-grammes	noms propres	étiquettes sémantiques	syno-nymes	mots reliés
151	WSJ870402-0112	41	34	67	69	525	91	1001	1001	148	60	38	49
151	WSJ870402-0118	58	35	17	32	1001	36	1001	1001	1001	100	42	1001
151	WSJ870413-0098	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
151	WSJ870414-0098	96	83	110	373	625	1001	900	1001	375	152	25	90

FIG. 3.9 – Exemple de matrice représentant pour chaque document pertinent (pour une requête) son rang dans la liste des résultats retournés par chaque index

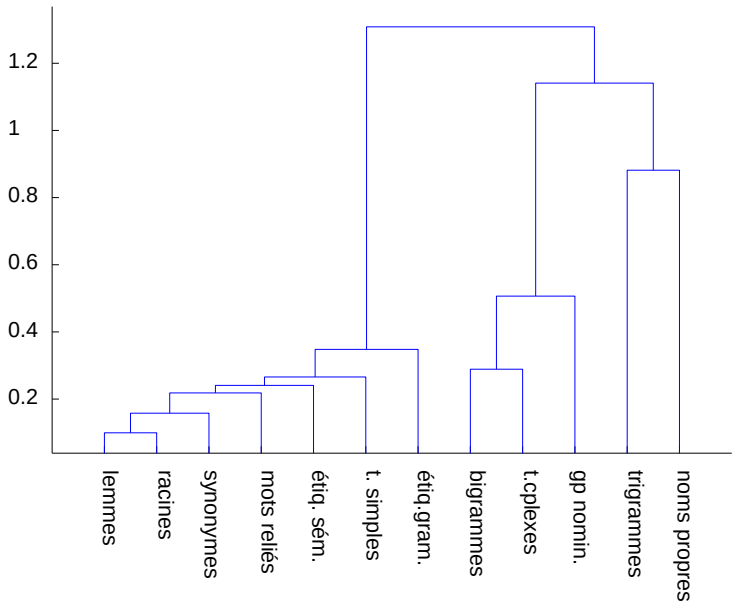

FIG. 3.10 – Classification ascendante hiérarchique obtenue à partir de la matrice des rangs des 12 index

À la suite de cette classification, deux groupes se dessinent nettement. Le premier réunit toutes les informations appartenant aux niveaux morphologique et sémantique de la langue. Au sein de ce groupe, les informations de lemmes, racines, synonymes, mots reliés et étiquettes sémantiques apparaissent particulièrement liées les unes autres (elles forment en effet une sous-classe distincte de celle des termes simples) ; nous avons vu précédemment que la relation qui les unit pouvait correspondre parfois à des cas de complémentarité pour retrouver les documents pertinents. Le second rassemble les informations syntaxiques (auxquelles viennent s'ajouter les noms propres, qui comme nous l'avons vu précédemment, ont un comportement qui diffère des autres connaissances sémantiques). Les informations de bigrammes et de termes complexes semblent également assez unies. Ces informations, nous l'avons dit, peuvent s'avérer complémentaires pour pallier les limites des méthodes d'acquisition sur lesquelles elles s'appuient (approche numérique *vs* symbolique).

Les résultats sont en accord avec les performances individuelles de chaque index puisque la première classe rassemble les informations linguistiques qui permettent au SRI de ramener le plus de documents pertinents, alors que la seconde regroupe les index qui ont le moins d'impact sur les performances des systèmes. Le fait que ce dendrogramme distingue nettement les informations syntaxiques des autres connaissances signifie que ces connaissances ont un comportement spécifique en RI. On ne peut pas pour autant en déduire que cette différence corresponde à un cas de complémentarité pour retrouver des documents pertinents puisque les informations syntaxiques, nous l'avons vu, ne contribuent généralement que très faiblement à l'amélioration des performances. Cette remarque est cependant étroitement liée aux types de connaissances syntaxiques et aux outils que nous avons exploités.

En conclusion, s'il existe des relations de complémentarité entre les informations linguistiques pour retrouver les documents pertinents, elles sont plutôt présentes soit à travers la combinaison d'informations morphologiques et sémantiques, soit au sein du couplage de connaissances mono-niveau (morphologique ou sémantique).

3.5 Bilan de la pertinence en RI du couplage d'informations multi-niveaux

Ce chapitre visait deux objectifs. Le premier était de mesurer dans un cadre homogène et sur des données identiques l'apport respectif en RI de différents types d'informations linguistiques, celles-ci étant généralement exploitées individuellement et dans des conditions diverses, rendant leur comparaison difficile. Le second était de se poser la question de la pertinence de leur couplage en RI et d'offrir les moyens d'y répondre. Pour cela, nous avons proposé de décrire les documents et requêtes selon 12 représentations linguistiques, en prenant en compte des informations standards, souvent utilisées dans les expériences combinant le TAL et la RI. Par le biais d'une plate-forme de test, ces diverses représentations ont été intégrées en parallèle au sein d'un SRI, ce qui nous a permis d'évaluer l'apport individuel des différentes informations linguistiques et d'analyser, par le biais d'une étude originale des corrélations, comment ces connaissances se comportaient les unes par rapport aux autres pour retrouver des documents pertinents.

Nos expérimentations conduisent à un certain nombre de remarques. Du point de vue de leur impact individuel sur les performances du SRI, seules les informations de niveaux morphologique (plus précisément les lemmes et les racines) et sémantique (en particulier les relations de synonymie et les liens morpho-sémantiques) se sont révélées véritablement efficaces pour améliorer les systèmes traditionnels. Contrairement à ceux souvent mitigés des travaux existants (présentés dans le chapitre précédent), nos résultats sont plus tranchés et attestent de l'intérêt de recourir à ces deux types de connaissances en RI. Les résultats obtenus en exploitant des connaissances syntaxiques sont toutefois décevants. Une des raisons pouvant justifier ces faibles performances est liée au fait que ces structures syntaxiques — bigrammes et trigrammes mis à part[12] —, bien que fortement significatives, sont peu nombreuses au sein des documents et requêtes de notre collection. La description des textes et des questions uniquement à partir de ce petit nombre de structures entraîne, comme nous l'avons vu, une sous-représentation des contenus textuels. Ce problème est dû d'une part au caractère généraliste de la collection utilisée et, d'autre part, aux outils utilisés pour leur extraction qui ne détectent pas toutes les structures syntaxiques présentes et ne gèrent pas, pour certains, leurs variations. C'est pourquoi, la plupart des travaux qui évaluent l'apport des connaissances syntaxiques en RI propose généralement, pour la représentation des documents et requêtes, de combiner ces connaissances à des termes simples au sein d'un même index. Bien que cette stratégie permette d'obtenir des résultats souvent meilleurs (*cf.* les expériences présentées en section 2.3.2.2 dans le chapitre précédent), elle ne correspond pas à nos objectifs puisque notre but, ici, était de considérer un seul type d'information linguistique par index.

Ces résultats obtenus au niveau individuel se retrouvent lorsque l'on combine les diverses connaissances et que l'on étudie les relations qu'elles entretiennent. Pour le couplage d'informations mono-niveau tout d'abord, il ressort des différentes analyses que ces connaissances, bien que souvent redondantes, peuvent dans certains cas être complémentaires pour aider le SRI à retrouver plus de documents que si elles étaient prises en compte individuellement. Ce constat s'applique néanmoins uniquement aux informations de nature morphologique ou sémantique. Les expérimentations effectuées n'ont pas permis en effet de démontrer l'intérêt en RI de combiner des informations de nature syntaxique (résultats bien sûr là encore fortement dépendants des outils et de la collection utilisés). Pour la prise en compte d'informations linguistiques multi-niveaux, il est apparu que seule la combinaison de connaissances bi-niveaux, principalement morpho-sémantique, pouvait présenter des cas intéressants de complémentarité. Là encore, leur couplage à des informations de type syntaxique ne paraît d'aucun intérêt pour retrouver davantage de documents pertinents.

Plus généralement, ces analyses qui ont mis en évidence l'apport significatif et tranché de certaines des informations exploitées et révélé des cas de complémentarité intéressants entre elles nous amènent à conclure à la pertinence du couplage d'informations linguistiques en RI. Puisque cet intérêt est montré, il reste à trouver la meilleure façon

[12] Pour les informations de bigrammes et de trigrammes, le problème est plus particulièrement lié au fait qu'aucun contrôle n'est effectué sur leur qualité. Elles ne sont pas, par conséquent, toujours très pertinentes pour refléter le contenu textuel des documents et requêtes.

de combiner ces informations dans un SRI afin d'exploiter de manière optimale leur efficacité, et ce, de manière automatique. C'est dans cette optique que nous avons envisagé une nouvelle technique qui propose de fusionner « intelligemment » les listes de résultats obtenues par chacune des informations linguistiques prises en compte. Cette technique de fusion, qui cherche à tirer le meilleur du couplage des informations linguistiques exploitées, fait l'objet du chapitre suivant.

Chapitre 4

Apprentissage pour la fusion de listes de résultats d'index linguistiques

Résumé : Afin de chercher à optimiser l'impact du couplage d'informations linguistiques multi-niveaux en RI, nous fusionnons les listes de résultats obtenues par chacun des index désignant une représentation linguistique particulière des documents et requêtes. Pour que cette fusion de données soit « intelligente », nous proposons une méthode originale qui, à partir d'un système d'apprentissage supervisé, est capable de déterminer, pour une requête donnée, la pertinence des documents de la collection en fonction d'une part de la position à laquelle ils ont été retrouvés au sein des listes des différents index et, d'autre part, d'informations caractérisant la requête. La liste de documents finalement détectés comme pertinents est alors évaluée et les résultats obtenus comparés à ceux d'un SRI déjà performant.

Mots-clés : fusion de listes de résultats, réseau de neurones, apprentissage artificiel supervisé, informations linguistiques multi-niveaux, caractéristiques des requêtes, rang des documents.

4.1 Introduction[1]

Dans le chapitre précédent, nous avons évalué l'intérêt d'utiliser et de coupler des informations linguistiques multi-niveaux en RI. Les diverses analyses proposées ont attesté de l'apport individuel de certaines des connaissances prises en compte — plus particulièrement d'ordre morphologique et sémantique — et ont révélé des cas de complémentarité intéressants. Ces résultats positifs — et même s'ils ne concernent qu'une partie de l'ensemble des informations linguistiques exploitées et révèlent des taux de

[1] Les travaux présentés dans ce chapitre ont été réalisés en étroite collaboration avec Vincent Claveau, chercheur au sein de l'équipe TexMex (IRISA), que nous remercions vivement pour son aide et son expertise en matière d'apprentissage.

complémentarité encore un peu faibles — permettent de valider la pertinence du couplage d'informations linguistiques en RI. Il reste maintenant à déterminer comment combiner efficacement et automatiquement ces connaissances au sein d'un SRI afin d'optimiser leur exploitation. Pour cela, nous proposons une méthode pour fusionner les listes de résultats (*i.e.* les listes de documents classés par ordre décroissant de pertinence par rapport à une requête donnée) obtenues par les divers index « linguistiques ». Cette fusion doit permettre d'obtenir une liste finale de résultats, correspondant aux meilleurs des documents retrouvés par les différents index, liste dont la pertinence peut alors être comparée à celle obtenue par un SRI traditionnel.

La fusion de ces listes est néanmoins problématique. Étant donné les résultats individuels obtenus par nos différents index, il n'est pas envisageable tout d'abord d'accorder la même importance à chacune de ces listes. Il est nécessaire en effet de privilégier les résultats des index qui se montrent efficaces pour améliorer les performances du SRI par rapport à ceux qui le sont moins. Nous ne souhaitons pas cependant donner *a priori* plus d'importance aux index qui semblent mieux se comporter que les autres. Leur efficacité peut en effet varier d'une requête à une autre et certains des index considérés globalement comme moins performants peuvent dans certains cas récupérer des documents intéressants. Nous voulons au contraire une technique de fusion qui soit souple et capable de s'auto-adapter. Enfin, une de nos hypothèses est que l'efficacité des index est tributaire du type de requête prise en compte et plus particulièrement de la nature des informations linguistiques qui la composent. En effet, il apparaît assez naturel que pour une requête qui contient par exemple un nom propre, les résultats obtenus par l'index des entités nommées puissent être privilégiés par rapport aux autres. Inversement, pour une requête composée uniquement de noms communs, ce type d'index a forcément moins d'impact. Il s'avère donc nécessaire d'adapter également l'importance à donner aux résultats des divers index en fonction des caractéristiques des requêtes prises en compte.

Toutes ces raisons justifient l'intérêt de concevoir une méthode pour la fusion de résultats qui soit souple et capable de détecter la meilleure façon de combiner les listes compte tenu d'une part des résultats individuels de chaque index et, d'autre part, du type de requêtes traitées. Pour cela, en nous appuyant sur une technique d'apprentissage supervisé, nous avons conçu un système, basé sur un réseau de neurones, qui, après une période d'apprentissage, est capable de déterminer, pour une requête donnée, la pertinence des documents en fonction d'une part de la position à laquelle ils ont été retrouvés au sein des listes des différents index et, d'autre part, d'informations caractérisant la requête. Il produit finalement une liste de résultats qui regroupe des documents détectés comme pertinents au sein des listes des différents index.

Ce chapitre présente donc le système d'apprentissage mis en place pour la tâche de fusion de listes et les résultats obtenus suite à son application. Nous nous intéressons dans un premier temps aux travaux proches de notre problématique (section 4.2). Nous présentons ensuite la méthode d'apprentissage supervisée proposée pour la fusion des résultats obtenus par un SRI exploitant différents index linguistiques (section 4.3). Nous décrivons ensuite les différentes expérimentations mises en œuvre pour son évaluation sur notre collection de documents (section 4.4) et discutons des résultats obtenus afin

de conclure (section 4.5) sur la pertinence en RI de fusionner les résultats d'un système qui combine diverses informations linguistiques.

4.2 Travaux connexes

Comme nous venons de l'évoquer, l'objectif de notre système est de fusionner les listes de résultats obtenues par un SRI qui intègre différentes représentations linguistiques des documents et requêtes, en nous appuyant notamment sur les caractéristiques des requêtes prises en compte. De nombreux travaux en RI se sont intéressés de manière indépendante à ces deux problématiques — la fusion des données et la prise en compte des spécificités des requêtes — pour améliorer les performances des systèmes. Avant de présenter plus en détail notre approche, qui a la particularité de coupler ces deux aspects, nous revenons dans cette section successivement sur les études qui s'intéressent d'une part aux problèmes de fusion de listes en RI (section 4.2.1) puis, d'autre part, à celles qui proposent de tenir compte des caractéristiques des requêtes, en particulier dans le but de prédire leur difficulté (section 4.2.2).

4.2.1 Fusion de données en RI

Le problème de la fusion de données, et plus particulièrement de celles des listes de résultats fournies par différents systèmes, a fait l'objet de nombreux travaux de recherche en RI (Croft, 1997; Beitzel *et al.*, 2004). Le point de départ de ces études est l'observation que certains SRI peuvent avoir des performances similaires sur une même collection de données (en termes de précision et de rappel) mais retrouver des documents pertinents différents. Puisque dans ce cas ces systèmes semblent complémentaires, la fusion de leurs résultats représente une idée intéressante. En effet, dans l'idéal, l'intersection des listes de résultats doit améliorer la précision des systèmes, puisque le fait qu'un document soit retrouvé par plusieurs systèmes atteste de sa pertinence ; l'union de ces listes doit favoriser le rappel puisque davantage de documents pertinents sont retrouvés après fusion des résultats des divers systèmes.

Les listes de résultats utilisées pour la fusion peuvent provenir de la combinaison de différentes sources d'information. Certains travaux (Belkin *et al.*, 1995; Fox et Shaw, 1994; Lee, 1997) proposent de multiplier les représentations d'une même requête afin de couvrir différentes facettes du besoin d'information de l'utilisateur. Il s'agit par exemple de combiner des requêtes de type booléen et des requêtes en langage naturel. D'autres s'intéressent au couplage de diverses stratégies de recherche, comme l'utilisation de différents modèles de RI, à travers par exemple la combinaison de modèles vectoriel et probabiliste (McCabe *et al.*, 1999) ou de plusieurs schémas de pondération (TF-IDF/BM25) (Savoy *et al.*, 1997; Lee, 1995). Des études couplent également divers types de descripteurs (obtenus par des indexations manuelle et automatique (Rajashekar et Croft, 1995), des descripteurs linguistiques (Perez-Carballo et Strzalkowski, 2000)...).

Quelle que soit la nature des données qui sont combinées, le problème demeure identique et consiste à déterminer comment fusionner les résultats de plusieurs recherches effectuées en parallèle afin d'obtenir de meilleures performances que si ces dernières

étaient appliquées de manière individuelle. Chaque type de recherche produit une liste ordonnée des documents retrouvés par le système et considérés comme pertinents par rapport à la requête de l'utilisateur. Chaque document est associé d'une part à un score de pertinence — déterminé par la fonction d'appariement du modèle de RI utilisé — qui représente son degré de similarité par rapport à la requête et, d'autre part, à un rang qui correspond à sa position dans le classement final (classement effectué dans l'ordre décroissant du score de pertinence obtenu par chaque document). Diverses techniques ont été proposées pour la fusion de ces listes et sont utilisées aussi bien pour des tâches de RI standard, de RI distribuée (où l'information est répartie entre plusieurs collections de documents) ou exploitées par des méta-systèmes de recherche (ou par des méta-moteurs dans le cas du *Web*). Nous n'en présentons ici que quelques-unes ; pour un état de l'art exhaustif, se référer à (Croft, 1997; Beitzel *et al.*, 2004). Ces techniques peuvent être différenciées selon qu'elles s'appuient sur les informations de rang des documents, sur leur score de pertinence et selon qu'elles nécessitent ou non des données d'entraînement.

Pour les approches qui exploitent essentiellement des informations de score de pertinence, divers algorithmes (Fox et Shaw, 1994; Lee, 1997) ont été proposés pour l'obtention d'un classement de résultats final. Ils consistent généralement, après avoir procédé à la normalisation des scores des documents présents dans chaque liste (valeur ramenée entre 0 et 1), à associer un score final à chaque document obtenu en additionnant chacun de ses scores individuels au sein des différentes listes. C'est notamment sur cette méthode que se fonde l'algorithme combSUM (Fox et Shaw, 1994) fréquemment utilisé dans des expériences de fusion. La liste de résultats finale correspond à l'ensemble des documents des différentes listes classés par ordre décroissant de leur nouveau score de pertinence. Plusieurs études ayant cherché à évaluer l'efficacité de ces approches ont montré que l'algorithme le plus efficace, combMNZ (Fox et Shaw, 1994; Bartell *et al.*, 1994; Vogt et Cottrell, 1999; Lee, 1997), consiste à faire la somme des scores normalisés d'un document qui a été retrouvé par diverses listes et à multiplier cette somme par le nombre de listes qui le contiennent. Il a été observé assez logiquement (Lee, 1997) que ces méthodes sont particulièrement performantes lorsque les différentes listes de résultats retrouvent des ensembles de documents pertinents très similaires et des ensembles très différents de documents non pertinents.

Il arrive cependant que les informations de score de pertinence ne soient pas toujours utilisables pour fusionner les listes de résultats, par exemple lorsque l'on compare divers modèles de RI qui utilisent des fonctions d'appariement très différentes ou bien lorsqu'un écart trop important entre les scores des documents des listes est observé. D'autres méthodes de fusion proposent donc d'exploiter les rangs des documents. Elles s'appuient généralement sur un système de vote (Aslam et Montague, 2001) où les documents sont les candidats et les différents systèmes (ou listes de résultats) sont les électeurs qui expriment leur préférence de rang sur les candidats. Plusieurs techniques de vote ont été utilisées pour la fusion de données en RI (Aslam et Montague, 2001). Nous pouvons par exemple citer la méthode de vote pondéré de Borda (1781) dont le principe peut être synthétisé comme suit :

– on dispose d'un ensemble de n documents (candidats) et de s systèmes (électeurs) ;

- chaque document issu de n reçoit des voix de la part des systèmes ;
- pour attribuer des voix à un document, un système s'appuie sur la position à laquelle il a retrouvé le document dans sa propre liste de résultats ; si le document a été classé dans cette liste en première position, il reçoit n voix, en deuxième position $n-1$... et seulement 1 voix si le document est le dernier de la liste ;
- le score final d'un document est la somme de toutes les voix qu'il a reçues de tous les systèmes.

Avec cette méthode, les documents qui sont privilégiés sont ceux qui ont été retrouvés comme les mieux classés dans les listes de résultats des différents systèmes. La technique de vote par majorité de Condorcet (1785) est également souvent utilisée. Bien que ces méthodes de vote semblent performantes en RI pour fusionner les résultats de différents systèmes (Renda et Straccia, 2003), plusieurs expériences (Mounir *et al.*, 1998) tendent à privilégier les approches basées sur les scores de pertinence.

Enfin, il est également important de préciser que les méthodes par rang ou par score ne sont pas incompatibles. Plusieurs travaux proposent de coupler ces deux types d'informations pour fusionner les listes de résultats (Perez-Carballo et Strzalkowski, 2000).

D'une manière générale, toutes ces méthodes sont non supervisées ; elles s'appuient en effet uniquement sur les informations retournées par les systèmes pour la fusion. Avec ce type d'approche, des documents non pertinents retrouvés dans plusieurs listes risquent d'avoir un poids important et d'être bien classés dans la liste finale. Plusieurs variantes s'appuyant sur des techniques supervisées ont été proposées pour pallier ces limites (*e.g.* les algorithmes *WeightedCondorcet-fuse* ou *WeightedCombMNZ*). Ces approches considèrent généralement le calcul d'un score de pertinence final d'un document comme une combinaison linéaire de ses scores normalisés au sein des différentes listes, elles-mêmes pondérées en fonction de leur efficacité à retrouver les documents pertinents. L'aspect supervisé de ce type d'approche intervient pour fixer l'importance qui est doit être donnée à chaque liste. Pour déterminer l'efficacité des différentes listes (et ainsi leur associer un poids), on s'appuie habituellement sur les performances individuelles de chaque système (comme la précision moyenne par exemple) dont l'évaluation nécessite de disposer de fichiers de pertinence (Bartell *et al.*, 1994). C'est notamment sur ce type d'approche que se base la méthode de fusion proposée par Strzalkowski *et al.* (1999).

Dans le cadre de nos travaux, ces méthodes de fusion de données sont problématiques puisqu'elles tendent à privilégier les documents qui sont retrouvés par un grand nombre de systèmes (ou présents dans un grand nombre de listes de résultats). Or, dans notre cas, un document peut être retrouvé par un seul des index exploités et être toutefois pertinent. Avec les techniques traditionnelles, l'importance qui sera donnée à ce document lors de la fusion sera par conséquent très faible. Cette idée justifie la nécessité de concevoir une méthode de fusion plus souple — qui permet notamment de s'appuyer sur d'autres critères que celui du nombre pour déterminer l'importance à donner aux documents présents dans les différentes listes —, capable d'exploiter le meilleur de chacune des listes de résultats prises en compte et de tirer parti des cas de complémentarité susceptibles d'exister entre elles.

Nous venons de présenter quelques-unes des méthodes de fusion de résultats utilisées en RI. Dans le système que nous proposons pour combiner nos listes de documents, nous utilisons des informations sur les rangs[2] des documents retrouvés mais également des connaissances sur les caractéristiques linguistiques des requêtes prises en compte. Avant de décrire plus en détail notre technique, nous présentons quelques travaux qui étudient ces connaissances sur les questions, dans le but de prédire leur difficulté et d'estimer ainsi la fiabilité des résultats obtenus par un SRI les traitant.

4.2.2 Prédiction de la difficulté de requêtes

Comme nous l'avons notamment évoqué dans le chapitre 2, l'une des conclusions qui ressort souvent des différentes expériences de RI est que la qualité des résultats des systèmes est fortement tributaire des requêtes traitées. Plusieurs travaux ont ainsi montré des variations de résultats importantes selon leur longueur (Voorhees et Harman, 1996)[3]. Récemment, des études ont cherché à mettre en évidence l'existence de relations entre les performances d'un SRI et les types de requêtes soumises, et ont observé une forte corrélation (Spärck Jones, 2000). Elles constatent plus précisément que certaines requêtes semblent plus difficiles que d'autres à traiter par un SRI, et que ce sont ces requêtes qui font chuter les performances du système. À partir de ces observations, l'hypothèse qui est faite est de considérer que si on arrive à prédire la difficulté des requêtes, il sera alors possible d'avoir une estimation de la qualité des résultats fournis par le système. La prédiction de la difficulté des requêtes offre donc de nouvelles perspectives intéressantes en RI. On peut ainsi par exemple imaginer des méthodes d'expansion de requêtes adaptées à leur niveau de difficulté (Mothe et Tanguy, 2005), ou envisager, pour les cas où les requêtes se révèlent trop complexes, des interactions avec l'utilisateur. En permettant une estimation de la fiabilité des résultats retournés par le SRI, la prédiction de difficulté peut également être utile aux méthodes de fusion de données (Yom-Tom *et al.*, 2005).

Plusieurs expériences ont été réalisées afin de déterminer, par le biais d'une analyse approfondie des requêtes, si certains des éléments qu'elles contiennent peuvent être des indices de leur difficulté. Pour évaluer si ces éléments sont de bons candidats, elles s'appuient généralement sur une étude de leurs corrélations avec les performances des systèmes (représentées par le biais de mesures traditionnelles en RI, comme la précision, le rappel, la MAP...). On peut distinguer les travaux qui étudient les caractéristiques linguistiques des requêtes de ceux qui prennent en compte d'autres paramètres généralement d'origine numérique (*e.g.* la fréquence des termes). Pour les premiers (Loupy et Bellot, 2000; Loupy, 2000; Mothe et Tanguy, 2005, *inter alia*), il a été observé que certaines informations linguistiques pouvant être extraites automatiquement des requêtes

[2]Le choix d'utiliser les rangs plutôt que les scores de pertinence est motivé par le fait que les performances obtenues par nos différents index (*cf.* section 3.4.2) sont très inégales ; les écarts importants observés entre les scores des documents d'une liste à une autre les rendent par conséquent difficilement exploitables.

[3]Ce phénomène avait notamment été observé au chapitre 2 lors de la description des expériences cherchant à évaluer l'impact d'une procédure de *stemming* sur les performances des SRI.

se révélaient être de bons indicateurs de leur difficulté. Ainsi, par exemple, dans les expériences de Mothe et Tanguy (2005), il est constaté que la présence de noms propres est une caractéristique fiable qui indique une requête « facile »[4], alors que ses complexités morphologique (notamment la présence de nombreux mots suffixés et composés de plusieurs morphèmes) et syntaxique (par exemple la distance syntaxique moyenne couverte par les relations syntaxiques) sont des indices révélant son caractère « difficile ». Dans ces travaux, l'ambiguïté sémantique des termes de la requête (obtenue en calculant le nombre moyen de sens de chaque mot à partir d'une ressource lexicale sémantique) apparaît également comme un bon indicateur de difficulté. Cette observation confirme les résultats obtenus par De Loupy et Bellot (Loupy et Bellot, 2000; Loupy, 2000). Toujours à un niveau sémantique, une autre caractéristique intéressante est la prise en compte du nombre d'hyponymes[5] que possède chaque terme de la requête (Jourlin *et al.*, 2000). Les informations d'hyponymie, obtenues par le biais de ressources sémantiques (*e.g.* WORDNET), permettent de mesurer la spécificité des termes de la requête. En effet, moins un terme a d'hyponymes, plus il est spécifique. Et, plus les requêtes contiennent de termes spécifiques, plus, dans ces travaux, elles auront de chances d'obtenir en réponse des documents pertinents. Enfin, d'autres caractéristiques linguistiques ont été évaluées (Grivolla, 2001), comme par exemple les informations de synonymie mais apparaissent toutefois des indicateurs moins efficaces.

Pour ce qui est des travaux cherchant à exploiter des indices de nature numérique présents au sein des requêtes, diverses expériences (Loupy et Bellot, 2000; Macdonald *et al.*, 2005, *inter alia*) confirment l'intérêt de prendre en compte la fréquence des termes d'une requête dans la collection. Il apparaît en effet qu'un mot présent dans moins de n documents de la collection peut être généralement considéré comme efficace pour la recherche. Cette observation est étroitement liée à l'idée, déjà évoquée précédemment, que plus les termes sont spécifiques, plus ils seront utiles pour la recherche. Diverses mesures ont ainsi été proposées pour calculer automatiquement la difficulté de la requête à partir de la fréquence moyenne de ses constituants dans la collection (Loupy et Bellot, 2000; Macdonald *et al.*, 2005; Bookstein *et al.*, 1995). Enfin, une mesure introduite par Cronen-Townsend *et al.* (2002) propose d'estimer le taux d'« ambiguïté »[6] de la requête (*clarity score*). Grâce à deux modèles de langue (l'un pour la requête, l'autre pour la collection de documents), il s'agit de déterminer si la requête est exprimée « dans le même langage » que celui utilisé au sein des documents de la collection. Une forte corrélation entre la précision moyenne des SRI et ce score a été constatée. Celui-ci est donc un indicateur fiable pour la prédiction de la difficulté de la requête.

L'ensemble de ces travaux atteste donc globalement de l'importance de l'analyse des requêtes en RI puisqu'un fort lien est observé entre les informations qui les caractérisent

[4]Une requête est considérée dans ce cas comme « facile » si une forte corrélation a pu être établie entre ses caractéristiques et les bonnes performances (en termes de précision et rappel) obtenues par le SRI. Inversement, une requête est « difficile » si une corrélation élevée a été observée entre ses caractéristiques et les faibles performances du SRI.

[5]L'hyponymie est la relation inverse de l'hyperonymie déjà définie en section 2.4.

[6]L'expression « taux d'ambiguïté » est la traduction proposée pour *clarity score* ; elle ne correspond pas exactement à la notion d'« ambiguïté » telle qu'elle est traditionnellement utilisée en linguistique.

et les performances des systèmes. Pour notre part, nous utilisons un certain nombre de caractéristiques linguistiques et numériques (informations de fréquence) dans le but, non pas de prédire la difficulté des requêtes, mais de mieux fusionner nos différentes listes de résultats. Notre idée est que ces caractéristiques peuvent être utiles à notre système pour apprendre, en s'appuyant également sur les informations de rang des documents, à identifier les documents qui sont pertinents pour une requête donnée au sein des différentes listes de résultats.

Le système d'apprentissage que nous proposons pour la fusion de listes de résultats, qui s'appuie donc à la fois sur des informations concernant la requête, sur les résultats obtenus par les différents index et sur une phase de supervision,

4.3 Système d'apprentissage supervisé pour la fusion de listes de résultats

Après avoir introduit quelques généralités sur l'apprentissage supervisé et présenté les principes généraux de la méthode d'apprentissage utilisée — les réseaux de neurones —, nous décrivons plus précisément l'approche retenue pour résoudre le problème de fusion de listes de résultats. L'apport de notre méthode ne se situe pas dans le domaine de l'apprentissage car nous ne sommes ici que des « utilisateurs » de techniques et outils existants. Notre objectif est de proposer une nouvelle approche, qui, par le biais d'un système d'apprentissage basé sur ces méthodes classiques, permet la fusion de listes de résultats à partir d'informations provenant, d'une part, de l'intégration de représentations linguistiques multi-niveaux au sein d'un SRI et, d'autre part, de caractéristiques issues de la requête à traiter.

4.3.1 Quelques généralités sur l'apprentissage supervisé

L'apprentissage artificiel (traduction de l'expression anglaise *machine learning*) dans une définition très générale consiste en l'élaboration de programmes qui s'améliorent avec l'expérience (Denis et Gilleron, 2000). Plus précisément, cette notion fait « référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances »[7] et « englobe toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement le modèle » (Cornuéjols et Miclet, 2002). Une application classique de l'apprentissage artificiel est la reconnaissance de caractères manuscrits, tels qu'ils apparaissent notamment sur une enveloppe. La difficulté réside principalement dans le nombre infini de formes différentes qu'il est possible de rencontrer. Ces diverses formes ne pouvant être recensées de manière exhaustive, il faut donc construire un système qui soit capable de généraliser la reconnaissance à partir d'un ensemble d'exemples de caractères.

[7] Cette définition est proposée l'AAAI (*American Association for Artificial Intelligence*) `http://www.aaai.org`.

Nous nous intéressons ici plus particulièrement aux techniques d'apprentissage supervisé, c'est-à-dire permettant d'apprendre automatiquement à partir d'exemples, qui s'appuient sur la notion d'induction. L'induction est le processus par lequel on tire des lois de portée générale en partant de l'observation de cas particuliers. Ces lois peuvent soit simplement permettre la prédiction sur une nouvelle observation, soit correspondre à une théorie générale du phénomène qui à la fois l'explique et permet de prédire.

Selon Cornuéjols et Miclet (2002), un problème d'apprentissage peut être défini de la manière suivante. Un système *apprenant* reçoit tout d'abord des données (des exemples) de l'univers dans lequel il est placé. Dans le cadre de l'apprentissage supervisé, chacune de ces données prend la forme d'un couple dans lequel on distingue d'une part la description d'une situation (ou *observation*), notée x_i (suivant une distribution D_X sur l'espace X de représentation des données) et, d'autre part, une réponse (ou *sortie désirée*), notée u_i, qui est fournie par un expert. Le système apprenant cherche alors à approximer au mieux la sortie désirée u_i pour chaque entrée observée x_i. Dans le cas idéal, le système est capable, après un certain temps d'apprentissage, de prédire exactement pour chaque entrée x_i, la réponse u_i. En apprentissage supervisé, on considère que pour associer à chaque forme x_i des couples-exemples une réponse u_i, l'expert, qui a étiqueté manuellement ces exemples, a utilisé une fonction inconnue par le système apprenant, appelée *fonction cible* et notée f (constante pour l'ensemble des exemples). Le système reçoit donc un échantillon d'exemples ou couples $(x_i, u_i) = (x_i, f(x_i))$, à partir duquel il doit chercher à deviner f ou au moins à en trouver une approximation h (souvent également nommée *classifieur*). Pour déterminer si le système doit continuer ou non son apprentissage, on s'appuie sur une fonction d'estimation de la qualité de h, généralement calculée sur des exemples. Cette description revient donc à voir l'apprentissage comme une tâche d'estimation de fonction — on cherche à apprendre un classifieur — à partir d'un échantillon de son comportement.

Nous venons de décrire de manière très succincte le principe inductif sur lequel s'appuie la plupart des techniques d'apprentissage artificiel supervisé. Il s'agit d'un objectif idéal. Il faut le distinguer de la méthode d'apprentissage, ou algorithme, qui décrit une réalisation effective de ce principe. Pour un principe inductif donné, il y a de nombreuses méthodes d'apprentissage possibles, résultant de choix différents pour régler les problèmes computationnels qui ne sont pas du ressort du principe inductif. Pour notre part, l'approche que nous proposons pour la fusion de listes de résultats s'appuie sur les réseaux de neurones. Ce choix est motivé par plusieurs raisons. D'un point de vue théorique tout d'abord, les réseaux de neurones sont particulièrement bien adaptés à traiter les données de nature exclusivement numérique que nous manipulons. D'un point de vue algorithmique, ces outils sont caractérisés par leur flexibilité et leur tolérance aux fautes et s'appuient sur des méthodes déjà bien établies. De plus, ils sont capables de manipuler un nombre important d'informations, ce qui, nous le verrons, est indispensable dans notre cas.

4.3.2 Réseaux de neurones : principes de base et apprentissage

L'objectif de cette section est d'une part de présenter de manière très rapide les principales propriétés des réseaux de neurones et, d'autre part, d'expliciter leurs capacités d'apprentissage. Une description plus générale de ces réseaux et de leurs applications peut notamment être trouvée dans (Cornuéjols et Miclet, 2002; Fiesler et Beale, 1996).

4.3.2.1 Principes

Dans les réseaux de neurones (ou réseaux connexionnistes), toutes les connaissances sont représentées par des liaisons entre les unités (neurones) et leurs poids synaptiques (valeurs) associés, par analogie avec la neurobiologie (Hertz *et al.*, 1991; Hérault et Jutten, 1994). Un réseau de neurones peut être représenté comme un graphe orienté et pondéré. Les nœuds de ce graphe sont des automates simples nommés neurones formels ou tout simplement unités du réseau. Un neurone formel possède plusieurs entrées d'information, est doté d'un état interne, que l'on appelle *état d'activation* (noté σ), et a une fonction de sortie f qui permet de calculer une valeur de sortie y en fonction de son état d'activation ($y = f(\sigma)$). Chaque neurone peut propager son état d'activation à une autre unité du réseau en passant par des arcs pondérés appelés connexions, liens ou poids synaptiques (notés w).

Un neurone formel est capable de faire seulement certaines opérations simples. Toute la puissance de calcul des réseaux de neurones réside dans l'interconnexion de ces unités élémentaires de traitement. Un réseau de neurones est donc caractérisé par son architecture, *i.e.* la structure selon laquelle les neurones qui le composent sont reliés les uns aux autres. Un exemple simple de réseau de neurones est le perceptron. Il est formé d'une couche de neurones d'entrée, *i.e.* des unités chargées de transmettre une composante du vecteur x (de l'espace X de représentation des données) et d'une couche de sortie représentée par un neurone unique qui fournit une hypothèse d'apprentissage — une approximation de f. Dans le cadre de la classification, il s'agit plus précisément d'une décision sur la classe à laquelle est attribuée x. Les connexions sont donc faites directement entre la couche d'entrée et le neurone de sortie. Dans ce type de réseau, la règle qui détermine la valeur du neurone de sortie (nommé *règle* ou *fonction d'activation*) correspond à une simple combinaison linéaire des valeurs d'entrée, et peut être représentée par la formule suivante : $y = f(\sigma) = w(0) + \sum_{i=1}^{d} w(i)x_i$ où $x_1, \ldots, x_i, \ldots, x_d$ représentent les données d'entrée du réseau (un ensemble de d d'attributs à valeurs dans $\{0, 1\}$ ou réelles), $w(0)$ est le poids d'un neurone formel (biais) ajouté automatiquement pour des raisons d'homogénéité de l'apprentissage (prenant toujours la valeur d'entrée 1), et $w(i)$ est le poids de la connexion entre le neurone d'entrée et le neurone de sortie. Un exemple de perceptron est présenté en figure 4.1.

Dans le cadre de nos travaux, nous utilisons un réseau de neurones multicouches, le modèle du PMC (perceptron multicouches), qui, en plus d'une couche de neurones d'entrée et d'une couche de sortie, possède également une ou plusieurs couches intermédiaires. Ces couches, nommées couches cachées, n'ont aucun contact avec l'environnement extérieur, leur rôle étant uniquement d'effectuer des calculs intermédiaires. Dans ce type d'architecture, les unités d'une couche sont reliées à toutes celles de la couche

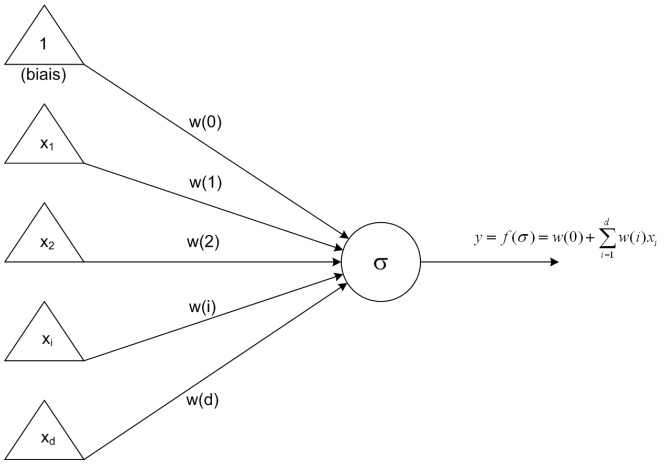

FIG. 4.1 – Exemple d'un réseau de neurone simple : le perceptron (figure inspirée de (Cornuéjols et Miclet, 2002))

suivante. Il n'existe aucune connexion entre les neurones d'une même couche (*cf.* figure 4.2). L'activation des réseaux de type perceptron multicouches est réalisée par la propagation des signaux à partir des unités d'entrée vers la sortie, en passant par les différentes couches cachées. D'une manière simplifiée, leur principe de fonctionnement est le suivant : la couche d'entrée est activée par l'arrivée d'une donnée, en recevant une composante du vecteur x sur chacun de ses neurones d'entrée. La première couche cachée effectue le calcul de l'état d'activation pour chacune de ces unités. Les valeurs de sortie sont transmises aux neurones de la couche cachée suivante, qui calculent à leur tour un nouvel état d'activation. Ce traitement est répété pour chacune des couches cachées. Finalement, l'unité de la couche de sortie ayant la valeur la plus forte indique la classe calculée pour l'entrée.

4.3.2.2 Apprentissage

La caractéristique la plus intéressante d'un réseau de neurones est sa capacité à apprendre, *i.e.* à modifier les poids de ses connexions en fonction des données d'apprentissage, de telle sorte qu'après une période d'entraînement, il ait acquis une faculté de généralisation. L'apprentissage est en général un processus graduel, itératif, où les poids du réseau sont modifiés plusieurs fois avant d'atteindre leurs valeurs finales. Nous nous plaçons ici dans un cadre d'apprentissage supervisé, ce qui signifie que nous disposons d'un comportement de référence précis — représenté par un ensemble d'exemples étiquetés par la classe à laquelle ils appartiennent — que nous désirons faire apprendre au réseau. Ce dernier doit être capable de mesurer la différence entre son comportement actuel et le comportement de référence, et de corriger ses poids de façon à réduire cette erreur.

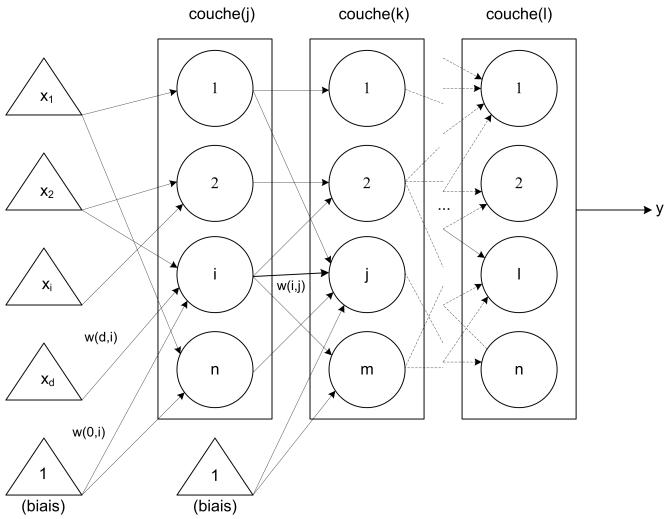

FIG. 4.2 – Exemple de perceptron multicouches (figure inspirée de (Cornuéjols et Miclet, 2002))

Dans un réseau de neurones de type perceptron, l'apprentissage s'effectue en modifiant les poids du réseau de façon à minimiser l'erreur entre la valeur obtenue par le neurone de sortie et la valeur attendue, telle qu'elle est donnée par l'exemple d'apprentissage. Il s'agit d'un algorithme basé sur la correction d'erreur. De manière simplifiée, l'apprentissage dans un perceptron peut être représenté par :

Initialiser les poids à l'aide de valeurs choisies au hasard ;
Répéter jusqu'à convergence
    Pour chaque exemple (couple (x_i, u_i))
        Calculer la valeur de sortie du réseau ;
        Tant que (valeur de sortie $\neq$ valeur de sortie désirée (u_i))
            Ajuster les poids ;
        Fin Tant que
    Fin Pour
Fin répéter jusqu'à

À la fin de l'apprentissage, lorsque le réseau a appris à modéliser son environnement, le comportement souhaité du réseau est le suivant : on présente un vecteur d'entrée au réseau ; celui-ci propage vers la sortie les valeurs d'activation correspondantes (en utilisant la règle de propagation) afin de générer par l'intermédiaire du neurone de sortie, une valeur de sortie. Celui-ci devrait correspondre à la sortie désirée, telle qu'elle a été apprise lors de la phase d'apprentissage.

L'apprentissage dans un réseau de type perceptron multicouches — utilisé dans notre système — est réalisé par une règle de rétro-propagation de l'erreur (Rumelhart *et al.*,

1986; Parker, 1985) qui permet d'entraîner d'une part les unités de sortie du réseau et, d'autre part, les unités cachées grâce à une technique de propagation en arrière de l'erreur à travers les couches du réseau.

Le principal avantage de l'apprentissage à l'aide de réseaux de neurones est d'être tolérant au bruit et aux erreurs. Le temps d'apprentissage peut être long ; en revanche, après apprentissage, le calcul de la sortie à partir d'un vecteur d'entrée est rapide. L'inconvénient le plus évident est que le résultat de l'apprentissage, c'est-à-dire le réseau de neurones calculé par l'algorithme d'apprentissage, n'est pas interprétable par l'utilisateur : on ne peut pas donner d'explication au calcul d'une sortie sur un vecteur d'entrée. On parle de « boîte noire ».

Nous venons de présenter quelques principes du fonctionnement des réseaux de neurones. Nous proposons à présent de voir comment nous avons utilisé cet algorithme d'apprentissage pour traiter notre problème de fusion de listes de résultats en RI.

4.3.3 Apprentissage supervisé pour la fusion de listes de résultats en RI

Le problème de fusion de listes de résultats tel que nous l'avons envisagé est abordé ici comme un problème de classification supervisée à deux classes. Il s'agit de déterminer si pour une requête donnée chacun des documents de la collection peut être considéré comme pertinent ou non pertinent compte tenu, d'une part, de son classement — pour la requête traitée — au sein des différentes listes de résultats et, d'autre part, d'informations caractérisant la requête. Tous les documents estimés pertinents feront alors partie de la liste de résultats finale.

La méthode proposée pour réaliser cette fusion se déroule en deux étapes : une phase d'apprentissage et une phase d'utilisation (*i.e.* la fusion proprement dite). Pour la phase d'apprentissage, il s'agit de présenter des couples requête-document à notre système qui, en s'appuyant sur des attributs associés à ce couple, c'est-à-dire le rang du document fourni par les différentes listes de résultats et des informations sur la requête (*cf.* section suivante), apprend, par le biais d'un réseau de neurones, à déterminer ce qui caractérise un document pertinent d'un document non pertinent pour la requête donnée. L'apprentissage est supervisé : pour cette étape, la pertinence du document est connue *a priori*[8].

À la suite de cette phase d'apprentissage, on obtient un classifieur qui, à partir des résultats des différentes listes et des caractéristiques des requêtes, est alors capable de distinguer les documents pertinents de ceux qui ne le sont pas. La phase d'utilisation consiste à présenter de nouveaux couples requête-document au classifieur qui détermine automatiquement la pertinence du document pour la requête. Les documents détectés comme pertinents sont alors retenus pour faire partie de la liste finale de résultats présentée à l'utilisateur.

Avant de décrire plus en détail l'architecture générale du système proposé, nous revenons tout d'abord sur les données d'entrée de notre système.

[8]Les informations de pertinence sont obtenues par les fichiers de TREC.

4.3.3.1 Données d'entrée

Les données d'entrée du système sont donc des couples requête-document — chaque document de la collection est évalué pour une requête donnée — représentés par un ensemble d'attributs. Ce sont sur ces attributs que notre approche se fonde pour déterminer, par le biais du système d'apprentissage, si un document peut être pertinent pour une requête et appartenir ainsi à la liste finale de résultats. Nous distinguons les attributs utilisés pour caractériser les documents de ceux qui concernent la requête.

Pour la représentation des documents tout d'abord, nous utilisons des informations directement issues des listes de résultats des différents index linguistiques exploités. Nous nous appuyons plus précisément sur les informations de rang associées aux documents retrouvés. Pour une requête donnée, on extrait pour chaque document la position à laquelle il a été classé au sein des différentes listes de résultats. Si le document est absent d'une liste, une valeur nulle lui est associée.

Pour la caractérisation des requêtes, nous nous appuyons sur les informations contenues dans la requête à traiter. Pour sélectionner ces caractéristiques, nous nous basons d'une part sur les travaux présentés précédemment sur la prédiction de la difficulté des requêtes et, d'autre part, sur une analyse manuelle de nos requêtes et de nos différents index afin de déterminer les connaissances susceptibles *a priori* d'être utiles pour retrouver les documents pertinents. À la suite de ces deux études, nous avons retenu une trentaine d'éléments de caractérisation des requêtes. Nous utilisons tout d'abord des informations permettant de prendre en compte l'influence de la longueur de la requête (nombre de phrases par requête, taille de la requête, nombre de mots pleins par requête). Nous nous appuyons également sur diverses informations linguistiques qu'elle contient et, plus précisément, sur des connaissances de nature morphologique (nombre de termes simples, de lemmes, de racines, de termes étiquetés grammaticalement par requête), syntaxique (nombre de verbes, de bigrammes, de groupes nominaux, de termes complexes, de trigrammes par requête) et sémantique (nombre de noms propres, de termes désambiguïsés, nombre moyen de sens et nombre de termes non ambigus par requête). Nous exploitons enfin des informations permettant de rendre compte de la spécificité d'une requête en s'appuyant sur la fréquence des termes qui la composent : fréquence de ces informations linguistiques dans la requête (fréquence moyenne des termes simples, des lemmes, des racines, des termes étiquetés grammaticalement, des bigrammes, des groupes nominaux, des termes complexes et des noms propres dans la requête) et dans les documents (fréquence documentaire moyenne des termes simples, des lemmes, des racines, des bigrammes, des groupes nominaux, des termes complexes et des noms propres de la requête).

Ces différents éléments sont extraits de manière automatique, soit directement à partir des informations contenues dans nos différents index, soit pour certaines informations sémantiques (*e.g.* le nombre de mots non ambigus) par le biais de ressources lexicales (WORDNET). Selon les requêtes, certaines de ces caractéristiques ne sont pas disponibles (toutes les requêtes ne contiennent pas, par exemple, de noms propres) ; une valeur nulle est alors attribuée.

L'ensemble de ces informations va constituer les différents attributs d'un couple requête-document présenté en entrée du système. Un tel couple peut donc être vu comme un vecteur (noté $\overrightarrow{req - doc}$) constitué de n composantes $x_1, \ldots x_i, \ldots, x_n$ qui correspondent à l'ensemble des attributs utilisés pour sa caractérisation. Ces données sont uniquement numériques. Le tableau 4.1 recense les caractéristiques en entrée du système représentant un document à classer pour une requête donnée.

caractérisation d'un document (pour une requête donnée)	
composante	description
x_1	rang du document retrouvé par l'index des termes simples
x_2	rang du document retrouvé par l'index des lemmes
x_3	rang du document retrouvé par l'index des racines
x_4	rang du document retrouvé par l'index des termes étiquetés grammaticalement
x_5	rang du document retrouvé par l'index des groupes nominaux
x_6	rang du document retrouvé par l'index des bigrammes
x_7	rang du document retrouvé par l'index des termes complexes
x_8	rang du document retrouvé par l'index des trigrammes
x_9	rang du document retrouvé par l'index des noms propres
x_{10}	rang du document retrouvé par l'index des termes étiquetés sémantiquement
x_{11}	rang du document retrouvé par l'index des synonymes
x_{12}	rang du document retrouvé par l'index des mots reliés morpho-sémantiquement
x_{13}	nombre de phrases par requête
x_{14}	taille de la requête (nombre de mots)
x_{15}	nombre de mots pleins par requête
x_{16}	nombre de termes simples par requête
x_{17}	nombre de lemmes par requête
x_{18}	nombre de racines par requête
x_{19}	nombre de termes étiquetés grammaticalement par requête
x_{20}	nombre de verbes par requête
x_{21}	nombre de bigrammes par requête
x_{22}	nombre de groupes nominaux par requête
x_{23}	nombre de termes complexes par requête
x_{24}	nombre de trigrammes par requête
x_{25}	nombre de noms propres par requête
x_{26}	nombre de termes désambiguïsés par requête
x_{27}	nombre moyen de sens par requête
x_{28}	nombre de termes non ambigus par requête
x_{29}	fréquence moyenne des termes simples dans la requête

composante	description
x_{30}	fréquence moyenne des lemmes dans la requête
x_{31}	fréquence moyenne des racines dans la requête
x_{32}	fréquence moyenne des termes étiquetés grammaticalement dans la requête
x_{33}	fréquence moyenne des bigrammes dans la requête
x_{34}	fréquence moyenne des groupes nominaux dans la requête
x_{35}	fréquence moyenne des termes complexes dans la requête
x_{36}	fréquence moyenne des noms propres dans la requête
x_{37}	fréquence documentaire moyenne des termes simples de la requête
x_{38}	fréquence documentaire moyenne des lemmes de la requête
x_{39}	fréquence documentaire moyenne des racines de la requête
x_{40}	fréquence documentaire moyenne des bigrammes de la requête
x_{41}	fréquence documentaire moyenne des groupes nominaux de la requête
x_{42}	fréquence documentaire moyenne des termes complexes de la requête
x_{43}	fréquence documentaire moyenne des noms propres de la requête
x_{44}	fréquence documentaire moyenne des termes simples de la requête

TAB. 4.1: Composantes d'un vecteur requête-document utilisées en entrée du système

Lors de la phase d'apprentissage, on associe à un couple requête-document la valeur de la classe attendue en sortie (pertinence ou non pertinence). Les exemples d'apprentissage sont donc de type $(\overrightarrow{req - doc}, \textit{décision de pertinence})$.

4.3.3.2 Architecture générale

Nous présentons maintenant l'organisation générale du système proposé pour la fusion de nos listes de résultats. La figure 4.3 illustre son architecture globale.

L'objectif final de notre approche est d'évaluer si la fusion des résultats obtenus par différents index linguistiques présente une valeur ajoutée par rapport aux performances d'un SRI exploitant un seul index déjà performant[9]. Comme nous l'avons déjà évoqué au début de cette section, la méthode de fusion proposée s'appuie sur un système qui fonctionne en deux temps. La première étape est la phase d'apprentissage proprement

[9]Comme nous le verrons plus loin, nous comparons nos résultats avec ceux obtenus par un SRI qui exploite l'index des racines qui avait obtenu sur notre collection la meilleure MAP dans les expérimentations du chapitre 3.

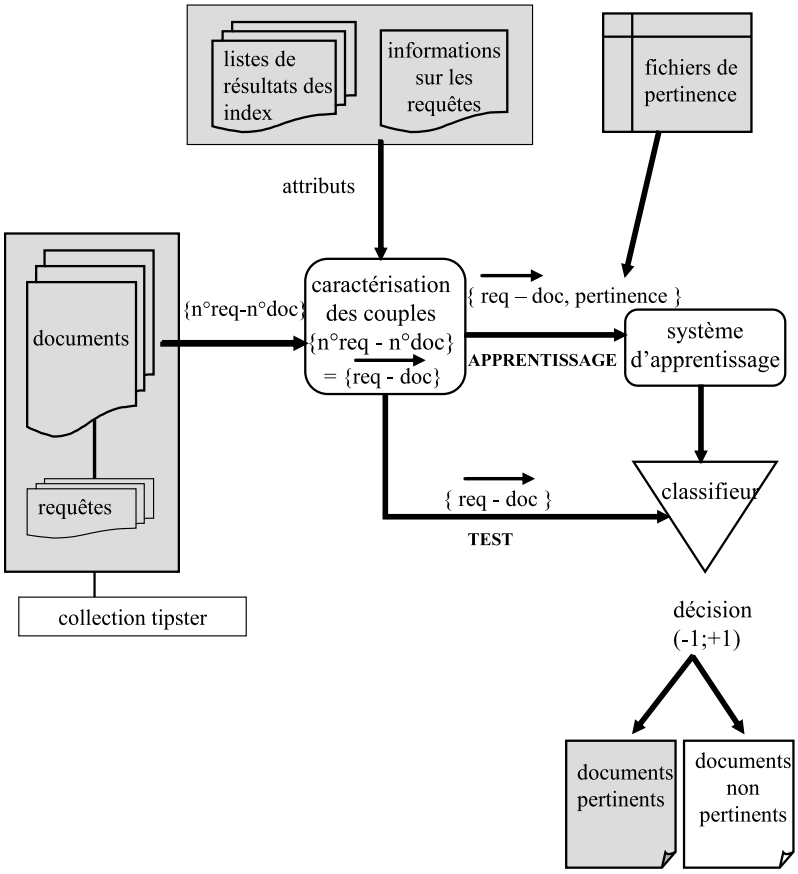

FIG. 4.3 – Architecture du système proposé pour la fusion des listes de résultats

dite qui consiste à présenter en entrée du réseau de neurones un ensemble de couples-exemples, qui correspondent dans notre cas à un vecteur requête-document caractérisé par un ensemble d'attributs, associés à leur décision de pertinence. À partir de ces exemples, le réseau de neurones apprend à distinguer les documents qui sont pertinents de ceux qui ne le sont pas pour une requête donnée (en fonction des éléments utilisés pour caractériser chaque couple requête-document). Lorsque le réseau a terminé son apprentissage, il devient un classifieur. Si l'apprentissage a été correctement effectué, le classifieur est alors capable de déterminer pour chaque document de la collection par rapport à une requête donnée sa pertinence compte tenu de sa position dans les différentes listes de résultats et en tenant compte des informations sur la requête. La deuxième étape consiste alors à tester les performances du classifieur et plus généralement à mesurer l'efficacité du système. Des couples requête-document sont présentés au classifieur qui fournit une décision sur leur pertinence. Les documents détectés comme pertinents pour une requête donnée sont regroupés dans une liste finale. Il convient d'évaluer alors si les documents contenus dans cette liste sont effectivement pertinents.

Cette évaluation revient à la fois à estimer la qualité du classifieur — on vérifie s'il a classé correctement ou non les documents — et à mesurer les performances de la méthode de fusion — la liste finale de résultats qui est évaluée correspond en effet à l'ensemble des documents de la collection qui ont été détectés comme pertinents (pour une requête) par le classifieur à partir des informations de rang associées aux documents et des caractéristiques de la requête. Nous insistons ici particulièrement sur le fait que nous réalisons une seule évaluation pour tester à la fois le classifieur et la méthode de fusion. Cette idée est fondamentale pour la compréhension de ce qui va suivre. Nous revenons à présent plus en détail sur le déroulement des phases d'apprentissage et de test.

4.3.3.3 Phase d'apprentissage

Constitution des couples-exemples

Il s'agit tout d'abord de constituer des couples d'exemples qui vont être ensuite utilisés comme entrées du système d'apprentissage proprement dit. Comme nous l'avons déjà mentionné, nous nous appuyons sur les informations issus des fichiers de pertinence de TREC. Nous extrayons donc pour chacune des 50 requêtes prises en compte, l'ensemble des documents de la collection et les jugements de pertinence qui leur sont associés[10]. Nous obtenons ainsi des exemples du type : ($n^o req - n^o doc$, *décision de pertinence*) où le couple $n^o req - n^o doc$ est composé des numéros d'identifiants de la requête *req* et du document *doc*, et *décision de pertinence* est une valeur binaire égale à 1 si le couple est pertinent, 0 sinon. Puisque notre méthode s'appuie sur les informations de rang et des connaissances sur la requête pour apprendre la pertinence d'un document, nous transformons chacun des couples requête-document de ces exemples par leur représentation sous forme d'attributs décrite précédemment. Nous obtenons donc des couples-exemples du type : ($\overrightarrow{req - doc}$, *décision de pertinence*) où le vecteur $\overrightarrow{req - doc}$ est représenté par un ensemble d'attributs $x_1, \ldots, x_{44}$ (*cf.* section 4.3.3.1). L'ensemble des couples ($\overrightarrow{req - doc}$, *décision de pertinence*) de la collection constituent donc nos exemples d'apprentissage.

Apprentissage et construction du classifieur

Pour chacun des exemples, $\overrightarrow{req - doc}$ est présenté en entrée du système d'apprentissage, et *décision de pertinence* correspond à la valeur de sortie désirée. À partir des données d'entrée et de la valeur de sortie attendue, le système — qui s'appuie sur un réseau de neurones de type perceptron multicouches —, apprend, par le biais de règles de modification des poids du réseau, un classifieur (fonction f) qui lui permet de généraliser ces exemples (*i.e.* d'expliquer le lien entre les données d'entrée et la classe de sortie). Une fois construit, le classifieur est capable de fournir une décision sur la

[10] Les documents de la collection non présents dans les fichiers de pertinence sont automatiquement considérés comme non pertinents.

pertinence d'un couple requête-document. Dans notre cas, il retourne plus précisément une valeur égale à -1 si le document est considéré comme non pertinent par rapport à une requête traitée, ou égale à 1 s'il est pertinent.

4.3.3.4 Phase de test

À la suite de cette phase d'apprentissage, il est nécessaire d'évaluer les performances du classifieur, *i.e.* sa capacité à classer correctement les documents pertinents ou non pertinents. Il s'avère toutefois indispensable de tester sa qualité sur d'autres données que celles utilisées pour l'apprentissage. En effet, il peut arriver que les hypothèses obtenues par le classifieur mémorisent l'ensemble des exemples d'apprentissage (*apprentissage par cœur*). Dans ce cas, il est estimé comme performant sur l'échantillon d'apprentissage mais aura un mauvais pouvoir de prédiction pour de nouvelles données.

Une méthode possible pour tester le classifieur est de partitionner l'échantillon de données disponibles (*i.e.* l'ensemble des couples requête-document et leurs jugements de pertinence) en un ensemble d'apprentissage (noté S) et un ensemble test (T). En procédant de la sorte, on peut alors supposer que les performances obtenues sur l'ensemble des données de test sont une bonne estimation de ses performances réelles. L'échantillon de données dont nous disposons est cependant assez petit (jeu de 50 requêtes) ; nous nous appuyons donc sur une autre méthode (la validation croisée) qui permet à la fois d'apprendre et de tester sur l'ensemble des données disponibles en obtenant malgré tout une estimation satisfaisante des performances du classifieur. Nous revenons plus en détail en section 4.4.2.1 sur la façon dont nous découpons nos données pour la phase d'apprentissage et la phase de test.

La phase de test consiste elle à présenter chaque couple requête-document au classifieur, qui fournit une décision de pertinence. Nous procédons de la sorte pour chacun des couples de l'échantillon de test — qui correspond, par le biais de la méthode la validation croisée, à l'ensemble des couples requête-document de la collection. Seuls les documents déterminés comme pertinents (pour une requête donnée) sont retenus dans la liste finale de résultats que nous produisons. Pour évaluer l'impact de notre méthode de fusion des résultats, nous comparons alors la pertinence de la liste de résultats obtenue à celle d'un SRI plus standard.

Nous venons de présenter l'organisation générale du système envisagé pour la fusion de listes de résultats issus de différents index linguistiques. Son efficacité est fortement liée à la qualité des données en entrée. Les attributs utilisés pour décrire les couples requête-document doivent être suffisamment discriminants et pertinents pour permettre au système de différencier les « bons » documents de ceux qui ne le sont pas. Nous faisons l'hypothèse ici que les informations de rang sur les documents retrouvés par les différents index et les diverses connaissances que nous avons sélectionnées pour décrire les requêtes sont des données suffisamment fiables pour permettre au système de réaliser efficacement sa tâche de fusion. Les performances de notre approche dépendent aussi de la qualité du système d'apprentissage mis en œuvre et de sa capacité, par le biais des

réseaux de neurones, à apprendre et généraliser les exemples — éventuellement bruités — pour construire un classifieur efficace.

4.4 Expérimentations et résultats

L'objectif de nos expérimentations est d'évaluer l'efficacité de notre méthode de fusion des résultats obtenus par divers index linguistiques. Il s'agit plus précisément, comme nous l'avons déjà dit, d'estimer la pertinence de la liste finale de documents produite en sortie de notre système et de la comparer à celle d'un SRI plus traditionnel.

Après avoir présenté brièvement les données nécessaires pour nos expérimentations (section 4.4.1), nous décrivons la méthodologie proposée pour le découpage des données nécessaires à la fois à l'apprentissage et au test du classifieur 4.4.2. Enfin, nous examinons et discutons les résultats obtenus par notre approche de fusion de listes sur notre collection de documents (section 4.4.3).

4.4.1 Description des données

Pour nos expérimentations, nous devons disposer de plusieurs données : une collection de test comme celles traditionnellement utilisées en RI — c'est-à-dire un ensemble de documents, de requêtes et de jugements de pertinence —, des listes de résultats fournis par nos différents index et obtenus à partir de cette même collection, et des informations sur les différentes requêtes.

Pour la collection de test, nous utilisons un sous-ensemble de la collection TIPSTER, déjà présenté dans le chapitre précédent, qui se compose d'un ensemble de 175000 documents (issus du *Wall Street Journal* des années 1986 à 1992), d'un jeu de 50 requêtes et de fichiers de pertinence (correspondant à la campagne de TREC-3 de 1994).

Les différentes listes de résultats sur lesquelles nous nous appuyons pour la fusion sont celles qui ont été obtenues au cours de nos expériences présentées au chapitre précédent. Nous utilisons donc les listes de documents retournés par le SRI (pour chacune des 50 requêtes de la collection) qui intègre 12 index différents, chacun correspondant à une représentation linguistique particulière des documents et requêtes. Pour rappel, les 12 informations linguistiques multi-niveaux exploitées sont les termes simples, les lemmes, les racines, les termes étiquetés grammaticalement, les bigrammes, les groupes nominaux, les termes complexes, les trigrammes, les noms propres, les termes étiquetés sémantiquement, les termes simples et leurs synonymes, et les termes simples associés à un ensemble de mots reliés morpho-sémantiquement. Chacune de ces listes de résultats contient l'ensemble des documents classés dans l'ordre décroissant de leur pertinence par rapport à la requête et retrouvés par le SRI exploitant un index particulier.

Enfin, les informations utilisées pour la description des requêtes (présentées en section 4.3.3.1), nous l'avons vu, sont extraites automatiquement à l'aide d'une part d'outils du TAL présentés également au chapitre précédent, de ressources construites *a priori* (pour les informations de nature sémantique), et de programmes conçus spécifiquement pour cette tâche (notamment pour extraire le nombre de mots non ambigus, le nombre moyen de sens par terme de la requête...).

Ainsi, à travers l'énumération de ces différentes données, nous constatons que notre système de fusion doit être capable de gérer un nombre considérable de données. Pour obtenir ses listes de résultats finales, il doit en effet prédire la pertinence de chaque document de la collection (soit environ 175000 documents au total) pour chacune des 50 requêtes en utilisant pour décrire chaque couple requête-document une quarantaine d'attributs. Le système que nous proposons pour la tâche d'apprentissage s'appuie sur un réseau de neurones de type perceptron multicouches, développé à partir de la librairie *open-source* FANN[11].

4.4.2 Méthodologie

Pour l'évaluation des performances du classifieur, il est nécessaire de s'appuyer, comme nous l'avons déjà mentionné, sur un jeu de test qui soit indépendant du jeu de données utilisé pour l'apprentissage. Nous présentons donc dans un premier temps la méthode sur laquelle nous nous basons pour le découpage des données. Nous décrivons ensuite les mesures d'évaluation retenues pour tester les performances du classifieur.

4.4.2.1 Découpage des données pour l'apprentissage et le test

L'estimation des performances du classifieur doit être effectuée sur un ensemble de données différentes de celles utilisées lors de la phase d'apprentissage. Étant donné le nombre limité de requêtes dont nous disposons pour cette évaluation (50), il est difficile d'envisager d'en sacrifier une partie pour tester le classifieur. Nous devons en effet prendre en compte l'ensemble des requêtes pour l'apprentissage si nous voulons construire un classifieur de bonne qualité. Pour contourner ce problème, nous nous appuyons sur la méthode de la validation croisée (illustrée en figure 4.4) qui permet d'utiliser le même jeu de données pour l'apprentissage et pour le test sans que cela affecte la qualité de l'évaluation. Cette méthode consiste plus précisément à partitionner aléatoirement l'ensemble des données en k sous-ensembles de même taille. On effectue alors k sessions d'apprentissage et k sessions de test de la façon suivante : pour une session, l'ensemble d'apprentissage est la réunion de $k-1$ sous-ensembles, et le test du classifieur s'effectue sur le $k^{ième}$ sous-ensemble restant. L'évaluation globale des performances du classifieur correspond alors à la moyenne des performances obtenues pour l'ensemble des k sessions de test (Denis et Gilleron, 2000). Dans notre cas, nous découpons notre échantillon de données — correspondant à l'ensemble des documents de la collection associés à leurs jugements de pertinence pour les 50 requêtes de notre collection — de manière aléatoire en 10 sous-ensembles de même taille. Chacun de ces sous-ensembles est donc composé de 5 requêtes et sert alternativement de jeu de test pour évaluer l'apprentissage effectué à partir des 9 autres sous-ensembles. La performance globale du classifieur est la moyenne de ses performances obtenues sur chacun des 10 jeux de test.

[11] La librairie FANN est disponible à l'adresse suivante : `http://leenissen.dk/fann/`.

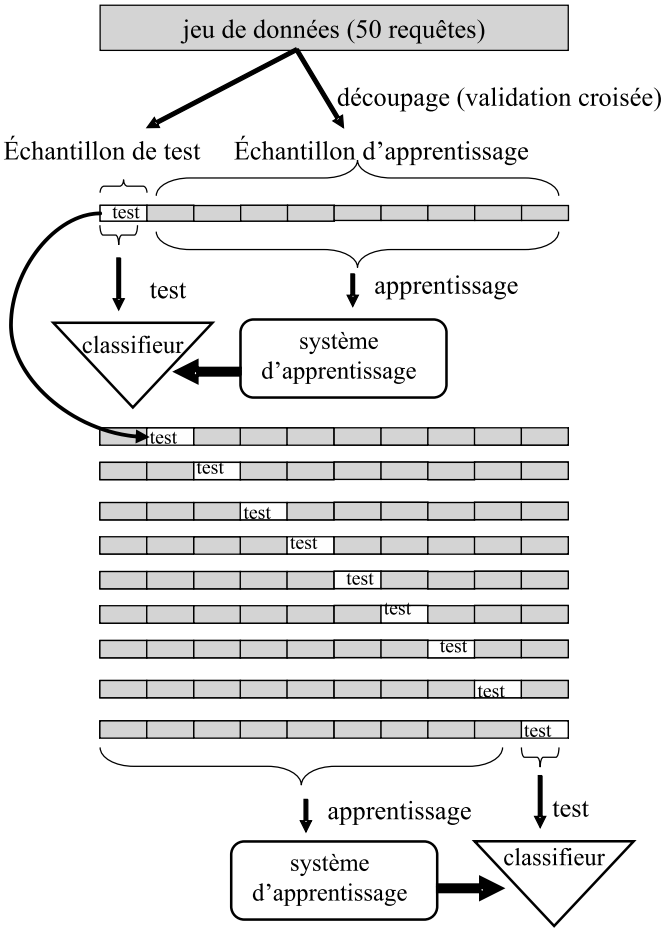

FIG. 4.4 – Illustration de la méthode de la validation croisée pour le découpage des données utilisées pour l'apprentissage et le test du classifieur

4.4.2.2 Mesures d'évaluation

Pour chaque jeu de test, nous devons évaluer si la décision de pertinence retournée par le classifieur pour chacun des documents retrouvés pour une requête donnée correspond à celle présente dans les fichiers de pertinence. Nous répétons cette opération pour les 5 requêtes issues d'un jeu de test. Pour cette évaluation, nous utilisons comme principale mesure la F-mesure (déjà décrite en section 1.4.2) traditionnellement employée en RI (avec un poids identique accordé à la précision et au rappel) et adaptée à l'évaluation de notre sortie de classifieur (binaire). Pour calculer la F-mesure, nous mesurons également les taux de rappel et de précision obtenus pour chaque requête.

Les listes de résultats évaluées (*i.e.* les documents qui ont été, pour une requête donnée, considérés par le classifieur comme pertinents) ne sont pas ordonnées ; seul un jugement binaire de pertinence associé aux documents pour chaque requête est établi.

Compte tenu de cette contrainte, nous ne pouvons donc utiliser les autres mesures habituelles, comme par exemple la précision moyenne non interpolée (MAP).

4.4.3 Résultats et discussions

L'objectif de nos expérimentations est d'évaluer l'efficacité de notre méthode qui, pour la fusion, s'appuie à la fois sur les rangs des documents et sur des informations sur les requêtes. Pour cela, nous procédons en plusieurs étapes. Nous évaluons dans un premier temps (section 4.4.3.1) la performance globale de notre classifieur (*i.e.* sa capacité à retrouver les documents pertinents) et nous la comparons à celle d'un SRI déjà efficace. Pour bien comprendre les résultats qui ont été observés, nous proposons ensuite (section 4.4.3.2) d'étudier plus en détail les performances obtenues sur chacune des requêtes évaluées. Enfin, une dernière expérience est réalisée (section 4.4.3.3) pour évaluer l'influence sur le fonctionnement et les résultats de notre méthode de la prise en compte d'informations sur la requête.

4.4.3.1 Évaluation globale de la méthode de fusion

À chacune des 5 requêtes d'un jeu de test est associée la liste (non ordonnée) des documents détectés par le classifieur comme pertinents. Pour évaluer la performance globale du classifieur et, par conséquent, de la pertinence des documents retrouvés, nous faisons la moyenne de la précision, du rappel et de la F-mesure calculés pour chacun des 10 jeux de tests évalués. Les résultats de notre méthode sont comparés à ceux obtenus sur la même collection par l'index observé comme le plus performant lors des expérimentations présentées au chapitre précédent : l'index des racines. Pour cette comparaison, nous souhaitons nous confronter au cas le plus difficile. Nous nous comparons donc aux meilleurs résultats obtenus par l'index des racines, évalués au DCV donnant la meilleure F-mesure (100 dans les expériences ci-après).

Le tableau 4.2 synthétise les moyennes des valeurs de précision, rappel et F-mesure obtenues sur l'ensemble des 10 jeux de tests pour les 2 systèmes, *i.e.* pour le SRI intégrant les racines et le système de fusion proposé. Les améliorations relatives (en %) observées apparaissent entre parenthèses.

	Index des racines (DCV=100)	Méthode de fusion (amélioration %)
Précision	29.70	29.48 (-0.73%)
Rappel	50.80	43.88 (-13.61%)
F-mesure	30.53	34.30 (+**12.33%**)

TAB. 4.2 – Moyennes des précision, rappel et F-mesure obtenues par la méthode de fusion et comparées aux performances d'un SRI exploitant des racines

Les nombres présentés au sein de ce tableau attestent de l'efficacité globale de notre méthode de fusion des index guidée par les informations sur les requêtes. Nous obtenons

en effet de bons résultats en termes de F-mesure puisqu'une amélioration relative de plus de 12% est observée par rapport aux performances obtenues par l'index des racines. En comparant l'efficacité de notre approche avec celle du SRI basé uniquement sur le meilleur index (les racines), nous montrons que notre méthode de fusion ne se contente pas d'exploiter les résultats de l'index le plus performant, mais qu'elle tire parti des autres index et des informations sur les requêtes pour proposer de meilleurs résultats.

Il ressort également de ce tableau que les résultats obtenus par notre approche sont, par rapport à ceux des racines, globalement identiques en termes de précision mais inférieurs pour le rappel. Le fait que nous obtenions toutefois une F-mesure plus élevée[12] laisse soupçonner que notre méthode offre des compromis rappel-précision plus équilibrés. Afin de vérifier cette hypothèse, nous proposons à présent d'étudier plus en détail les performances des 2 systèmes pour chacune des requêtes traitées.

4.4.3.2 Analyse des performances requête par requête

En analysant manuellement les résultats des 2 systèmes requête par requête, nous constatons que les performances observées pour l'index de racines varient fortement d'une requête à l'autre. Pour certaines, il obtient de très bons résultats en termes de rappel — proches des 100% — généralement associés à une précision très faible, proche des 5%, ce qui lui permet d'avoir un rappel moyen plus élevé que le nôtre mais une F-mesure très faible (environ 10). De manière générale, les résultats obtenus par cet index semblent plus irréguliers que ceux obtenus par notre méthode. Afin de confirmer expérimentalement cette variation des résultats, nous proposons de calculer, pour notre méthode et pour le SRI qui exploite les racines, la moyenne et l'écart-type des valeurs de précision, de rappel et F-mesure obtenues pour chacune des 5 requêtes d'un jeu de test[13]. L'écart-type nous permet ainsi d'avoir une indication de la dispersion des résultats observés pour les diverses requêtes par rapport aux moyennes considérées (moyennes calculées sur un jeu de test). Les tableaux suivants présentent les résultats sur chacun des 10 jeux de test, pour la F-mesure (tableau 4.3), la précision (tableau 4.4) et le rappel (tableau 4.5).

Les différences observées entre les résultats des deux systèmes pour les valeurs d'écart-type sont particulièrement significatives, et ce, pour les trois mesures. Pour notre méthode, nous constatons, sur les 10 jeux de test évalués, un écart-type très faible. En effet, les valeurs observées sont comprises entre 1.64 et 5.82 pour la F-mesure, entre 1.96 et 4.38 pour la précision et entre 2.44 et 9.70 pour le rappel, ce qui signifie que les résultats obtenus pour chacune des requêtes sont très proches de la moyenne. Les résultats de notre approche sont donc quasi constants quelle que soit la requête traitée. Pour le SRI basé sur les racines, l'écart-type est nettement plus élevé. Les différentes valeurs obtenues sont comprises entre 11.26 et 20.55 pour la F-mesure, entre 11.73 et 26.15 pour

[12] Nous rappelons que, selon le principe même de la moyenne harmonique, une F-mesure faible est obtenue lorsque l'un des deux taux (précision ou rappel) est bas même si l'autre est très élevé ; inversement, elle est forte si les deux taux sont proches et élevés.

[13] Bien que le jeu de requêtes de la collection TIPSTER contiennent 50 requêtes, seules 49 sont effectivement évaluées dans les fichiers de pertinence issus de TREC. Cela explique que sur les 10 jeux de test utilisés pour l'évaluation, le dernier ensemble ne contient que 4 requêtes.

Découpage des données de test (validation croisée)	Requêtes (numéros)	Racines (DVC=100)	Fusion
		Moyenne F-mesure (écart-type)	Moyenne F-mesure (écart-type)
Jeu de test 1	(169-173-180-183-196)	45.64 (13.32)	46.55 (5.61)
Jeu de test 2	(153-156-190-199-200)	23.33 (11.26)	24.81 (5.82)
Jeu de test 3	(155-164-166-168-191)	30.43 (16.67)	36.72 (4.42)
Jeu de test 4	(154-162-167-172-178)	21.68 (18.32)	32.51 (3.33)
Jeu de test 5	(181-185-186-187-198)	28.13 (18.32)	40.49 (4.25)
Jeu de test 6	(158-170-171-179-182)	33.34 (15.74)	34.17 (4.04)
Jeu de test 7	(159-161-192-194-195)	29.76 (20.55)	30.25 (3.29)
Jeu de test 8	(160-165-184-193-197)	26.79 (11.87)	27.40 (2.12)
Jeu de test 9	(152-157-163-174-175)	40.67 (17.20)	43.98 (1.64)
Jeu de test 10	(151-176-177-189)	25.54 (17.68)	26.10 (4.70)
Moyenne sur les 10 jeux de test		30.53 (16.09)	34.30 (3.92)

TAB. 4.3 – Moyennes et écarts-type des valeurs de **F-mesure** (calculés pour 5 requêtes) obtenus sur les 10 jeux de test par l'index des racines et par la méthode de fusion

Découpage des données de test (validation croisée)	Requêtes (numéros)	Racines (DVC=100)	Fusion
		Moyenne précision (écart-type)	Moyenne précision (écart-type)
Jeu de test 1	(169-173-180-183-196)	49.59 (24.49)	47.47 (4.37)
Jeu de test 2	(153-156-190-199-200)	18.11 (11.73)	18.28 (4.38)
Jeu de test 3	(155-164-166-168-191)	23.54 (15.30)	28.37 (3.71)
Jeu de test 4	(154-162-167-172-178)	19.95 (21.15)	25.23 (3.97)
Jeu de test 5	(181-185-186-187-198)	43.30 (22.92)	36.52 (4.09)
Jeu de test 6	(158-170-171-179-182)	26.59 (14.13)	25.98 (4.17)
Jeu de test 7	(159-161-192-194-195)	23.67 (18.07)	24.16 (3.41)
Jeu de test 8	(160-165-184-193-197)	22.07 (13.33)	19.08 (1.96)
Jeu de test 9	(152-157-163-174-175)	37.62 (25.93)	40.16 (3.83)
Jeu de test 10	(151-176-177-189)	32.31 (26.15)	29.51 (4.26)
Moyenne sur les 10 jeux de test		29.70 (19.32)	29.48 (3.81)

TAB. 4.4 – Moyennes et écarts-type des valeurs de **précision** (calculés pour 5 requêtes) obtenus sur les 10 jeux de test par l'index des racines et par la méthode de fusion

la précision et entre 7.85 et 32.49 pour le rappel, traduisant ainsi une dispersion nettement plus importante des résultats des différentes requêtes par rapport à la moyenne. L'index des racines apporte des améliorations très significatives pour certaines requêtes

Découpage des données de test (validation croisée)	Requêtes (numéros)	Racines (DVC=100)	Fusion
		Moyenne rappel (écart-type)	Moyenne rappel (écart-type)
Jeu de test 1	(169-173-180-183-196)	47.57 (13.41)	43.18 (7.16)
Jeu de test 2	(153-156-190-199-200)	47.13 (13.65)	38.86 (9.70)
Jeu de test 3	(155-164-166-168-191)	61.21 (11.32)	52.33 (6.89)
Jeu de test 4	(154-162-167-172-178)	55.32 (32.49)	46.51 (2.44)
Jeu de test 5	(181-185-186-187-198)	40.31 (18.60)	45.52 (5.11)
Jeu de test 6	(158-170-171-179-182)	52.57 (26.11)	50.93 (7.68)
Jeu de test 7	(159-161-192-194-195)	49.03 (29.66)	40.75 (3.36)
Jeu de test 8	(160-165-184-193-197)	41.73 (7.85)	47.57 (3.95)
Jeu de test 9	(152-157-163-174-175)	75.53 (25.99)	49.50 (6.07)
Jeu de test 10	(151-176-177-189)	37.59 (21.10)	23.63 (5.36)
Moyenne sur les 10 jeux de test		50.80 (20.02)	43.88 (5.77)

TAB. 4.5 – Moyennes et écarts-type des valeurs de **rappel** (calculés pour 5 requêtes) obtenus sur les 10 jeux de test par l'index des racines et par la méthode de fusion

mais s'avère moins efficace pour d'autres en proposant des résultats déséquilibrés (fort rappel et précision très basse ou le contraire).

Le fait que les résultats obtenus soient plus stables témoigne de la capacité de notre méthode à compenser les cas, où pour une requête donnée l'index des racines échoue, en s'appuyant sur les résultats des autres index. Le principal apport de notre méthode est donc d'effectuer un lissage des résultats obtenus par les différents index et de les rendre par conséquent moins sensibles aux types de requêtes prises en compte.

Puisque les résultats observés sont constants sur les différentes questions évaluées, on peut émettre l'hypothèse que le système s'est adapté aux spécificités des requêtes pour sélectionner les index susceptibles d'être les plus efficaces pour retrouver les documents pertinents. Pour confirmer cette idée, nous proposons une nouvelle expérimentation permettant d'évaluer l'influence sur l'efficacité de notre méthode de la prise en compte d'informations sur les requêtes.

4.4.3.3 Influence des caractéristiques des requêtes sur l'efficacité de notre méthode de fusion

Un moyen simple de valider expérimentalement l'hypothèse que notre méthode de fusion s'appuie sur les caractéristiques des requêtes pour sélectionner les meilleurs index est de réitérer les expériences précédentes en enlevant de notre système l'ensemble des attributs utilisés qui correspondent à des informations sur la requête (*i.e.* les attributs $x_{13}, \ldots, x_{44}$). Cette suppression revient à faire apprendre notre réseau de neurones uniquement à partir des informations de rang sur les documents. En comparant les

résultats obtenus lors des deux expériences — *i.e.* les performances avec et sans prise en compte d'informations sur la requête — nous pourrons avoir une idée précise du fonctionnement de notre méthode de fusion et, plus particulièrement, de la façon dont le réseau de neurones exploite les attributs associés aux couples requête-document pour apprendre le classifieur. Ces expérimentations permettent de compenser partiellement le fonctionnement en « boîte noire » des réseaux de neurones qui n'offrent aucune interprétation des résultats obtenus.

Le tableau (4.6) synthétise, comme précédemment, les moyennes des valeurs de précision, rappel et F-mesure obtenues sur l'ensemble des 10 jeux de test pour le SRI intégrant les racines et pour notre méthode de fusion à laquelle nous avons retiré les informations sur les requêtes. Les améliorations relatives (en %) observées apparaissent entre parenthèses.

	Index des racines	Fusion sans infos req. (amélioration %)
Précision	27.90	26.49 (-5.04%)
Rappel	51.32	41.65 (-18.84%)
F-mesure	29.67	30.28 (+2.05%)

TAB. 4.6 – Moyennes des précision, rappel et F-mesure obtenues par la méthode de fusion sans prise en compte des informations sur la requête et comparées aux performances d'un SRI exploitant des racines

Les tableaux (tableau 4.7, tableau 4.8 et 4.9) présentent les résultats respectivement obtenus sur chacun des 10 jeux de test pour la F-mesure, la précision et le rappel.

Découpage des données de test (validation croisée)	Requêtes (numéros)	Racines	Fusion sans infos req.
		Moyenne F-mesure (écart-type)	Moyenne F-mesure (écart-type)
Jeu de test 1	(151-158-178-194-199)	17.15 (15.60)	19.89 (4.09)
Jeu de test 2	(153-174-183-196-198)	45.23 (17.06)	46.67 (4.50)
Jeu de test 3	(159-160-169-184-195)	23.65 (14.42)	24.58 (3.16)
Jeu de test 4	(152-155-163-164-197)	38.42 (11.71)	37.18 (4.89)
Jeu de test 5	(154-156-179-187-193)	26.55 (16.04)	24.37 (1.74)
Jeu de test 6	(168-175-182-190-200)	30.32 (12.66)	22.65 (3.07)
Jeu de test 7	(171-172-180-181-185)	21.58 (13.88)	24.14 (4.73)
Jeu de test 8	(157-165-173-176-188)	31.21 (20.90)	42.62 (8.57)
Jeu de test 9	(162-167-186-191-192)	27.52 (13.46)	28.32 (1.90)
Jeu de test 10	(166-170-177-189)	35.07 (19.92)	32.38 (3.03)
Moyenne sur les 10 jeux de test		29.67 (15.57)	30.28 (3.97)

TAB. 4.7 – Moyennes et écarts-type des valeurs de **F-mesure** (calculés pour 5 requêtes) obtenus sur les 10 jeux de test par l'index des racines et par la méthode de fusion sans prise en compte des informations sur les requêtes

Découpage des données de test (validation croisée)	Requêtes (numéros)	Racines	Fusion sans infos req.
		Moyenne précision (écart-type)	Moyenne précision (écart-type)
Jeu de test 1	(151-158-178-194-199)	13.73 (16.99)	12.53 (3.02)
Jeu de test 2	(153-174-183-196-198)	45.91 (20.07)	41.07 (6.02)
Jeu de test 3	(159-160-169-184-195)	18.41 (9.79)	17.78 (3.29)
Jeu de test 4	(152-155-163-164-197)	36.70 (22.10)	33.50 (4.73)
Jeu de test 5	(154-156-179-187-193)	35.05 (26.06)	35.18 (4.69)
Jeu de test 6	(168-175-182-190-200)	22.34 (11.24)	15.62 (2.22)
Jeu de test 7	(171-172-180-181-185)	16.53 (13.85)	15.99 (3.68)
Jeu de test 8	(157-165-173-176-188)	33.98 (30.41)	39.69 (7.14)
Jeu de test 9	(162-167-186-191-192)	21.90 (11.64)	21.74 (1.75)
Jeu de test 10	(166-170-177-189)	34.48 (25.90)	31.84 (3.49)
Moyenne sur les 10 jeux de test		27.90 (18.81)	26.49 (4)

TAB. 4.8 – Moyennes et écarts-type des valeurs de **précision** (calculés pour 5 requêtes) obtenus sur les 10 jeux de test par l'index des racines et par la méthode de fusion sans prise en compte des informations sur les requêtes

Découpage des données de test (validation croisée)	Requêtes (numéros)	Racines	Fusion sans infos req.
		Moyenne rappel (écart-type)	Moyenne rappel (écart-type)
Jeu de test 1	(151-158-178-194-199)	60.83 (24.23)	49.30 (5.84)
Jeu de test 2	(153-174-183-196-198)	50.78 (11.51)	54.41 (2.52)
Jeu de test 3	(159-160-169-184-195)	40.11 (25.12)	40.53 (3.24)
Jeu de test 4	(152-155-163-164-197)	57.38 (24.45)	42.12 (6.84)
Jeu de test 5	(154-156-179-187-193)	30.13 (15.46)	18.73 (1.09)
Jeu de test 6	(168-175-182-190-200)	62.45 (21.99)	41.70 (6.64)
Jeu de test 7	(171-172-180-181-185)	55.19 (14.44)	49.94 (5.73)
Jeu de test 8	(157-165-173-176-188)	46.87 (26.31)	46.10 (10.53)
Jeu de test 9	(162-167-186-191-192)	47.79 (22.77)	40.68 (1.84)
Jeu de test 10	(166-170-177-189)	61.69 (27.03)	32.98 (2.88)
Moyenne sur les 10 jeux de test		51.32 (21.33)	41.65 (4.72)

TAB. 4.9 – Moyennes et écarts-type des valeurs de **rappel** (calculés pour 5 requêtes) obtenus sur les 10 jeux de test par l'index des racines et par la méthode de fusion sans prise en compte des informations sur les requêtes

Le tableau 4.10 synthétise les résultats de notre méthode de fusion avec et sans prise en compte des informations sur la requête.

	Fusion infos req. + rang Moyenne (écart-type)	Fusion sans infos req. Moyenne (écart-type)
Précision	29.48 (3.81)	26.49 (4)
Rappel	43.88 (5.77)	41.65 (4.72)
F-mesure	34.30 (3.92)	30.28 (3.97)

TAB. 4.10 – Moyennes et écarts-type des valeurs de précision, rappel et F-mesure obtenus sur les 10 jeux de test par la méthode de fusion avec et sans prise en compte des informations sur les requêtes

À partir des nombres présentés dans ces différents tableaux, nous pouvons constater que lorsque notre méthode n'exploite aucune information sur les requêtes, les résultats sont nettement moins bons que ceux obtenus précédemment. Ces observations mettent en avant l'idée que notre méthode s'appuie en grande partie, pour la combinaison des résultats, sur les requêtes pour identifier les index susceptibles d'être les plus efficaces pour retrouver des documents pertinents. En ce sens, elle est plus complexe que les techniques de fusion traditionnelles basées uniquement sur les listes de résultats, mais

aussi plus souple puisqu'elle est capable de s'adapter différemment à chaque type de requêtes.

Plus généralement, compte tenu des résultats obtenus à la suite de nos différentes expérimentations, nous pouvons conclure que la combinaison d'informations linguistiques multi-niveaux constitue une piste intéressante à explorer pour le couplage TAL-RI. En effet, comme nous l'avons évoqué en introduction générale et lors de la synthèse proposée dans le chapitre 2, un des points faibles constatés dans la plupart des travaux cherchant à évaluer l'apport d'une connaissance linguistique particulière au sein d'un SRI est le caractère très irrégulier des résultats obtenus, souvent tributaires de nombreux facteurs (comme le type de collection, la longueur des requêtes, des documents...). En combinant les potentialités de plusieurs informations linguistiques au sein d'un même système, comme nous l'avons proposé ici par le biais de notre méthode de fusion, nous montrons qu'il est possible d'atténuer la variation des résultats. Pour cela, la méthode que nous avons conçue apparaît particulièrement efficace puisqu'elle est suffisamment souple pour s'adapter aux divers types de requêtes et optimiser l'exploitation des différentes informations linguistiques.

4.5 Conclusion

L'objectif de ce chapitre était d'évaluer l'impact du couplage d'informations linguistiques en RI. Pour cela, il était nécessaire de trouver une méthode qui permette de combiner efficacement ces connaissances afin d'optimiser leur exploitation. Nous avons donc proposé de fusionner les listes de résultats obtenus par un SRI intégrant diverses représentations linguistiques des documents et requêtes. Pour mener à bien cette tâche de fusion, nous avons développé une méthode qui, par le biais d'un système d'apprentissage supervisé basé sur un réseau de neurones, est capable d'identifier la pertinence des documents de la collection en fonction de leur position dans les différentes listes de résultats et à partir d'informations sur la requête. L'ensemble des documents détectés comme pertinents sont regroupés dans une liste finale dont la pertinence a été évaluée.

Les résultats obtenus montrent un gain de notre méthode de fusion, non pas en termes de précision et de rappel comme nous aurions pu nous y attendre, mais dans le meilleur compromis qu'elle offre de ces deux taux à chaque requête. Ils attestent également de la capacité de notre méthode de fusion à s'adapter différemment à chaque type de requêtes. La stabilité des résultats observés montre que l'exploitation conjointe de plusieurs informations linguistiques permet de compenser les faiblesses des connaissances prises en compte individuellement, liées à leur variation de résultats selon les types de données manipulées et, plus particulièrement dans notre cas, selon les requêtes considérées.

Ces expérimentations constituent une première piste intéressante pour mettre en évidence l'idée qu'en couplant autrement le TAL et la RI, les résultats obtenus peuvent être plus tranchés. Le fait que notre classifieur ne retourne qu'une décision binaire sur la pertinence des documents est néanmoins problématique puisqu'il ne nous permet pas de comparer plus efficacement les performances de notre méthode à celles obtenues

par d'autres SRI. L'aspect « boîte noire » des réseaux de neurones sur lesquels s'appuie notre méthode rend également difficile la compréhension des processus mis en œuvre.

Toujours dans le but de démontrer l'intérêt en RI de recourir à des informations linguistiques, à condition que ces dernières soient exploitées adéquatement, nous nous intéressons dans le dernier chapitre de ce mémoire à un type de connaissance précis qui s'est révélé tout au long de nos expérimentations le plus efficace pour améliorer les performances des SRI : les informations morphologiques. Bien que d'autres travaux aient déjà démontré leur apport, nous avons pu constater, aussi bien lors de la synthèse de leur exploitation en RI proposée au chapitre 2, que lors de l'évaluation du SRI basé sur l'index des racines que nous venons de présenter, que les améliorations obtenues suite à leur prise en compte étaient fortement fluctuantes et dépendantes des spécificités des données manipulées (taille des documents, longueur des requêtes, langue prise en compte...). Étant donné l'impact de ces informations sur les performances des systèmes, il nous semble intéressant d'envisager une nouvelle approche visant à obtenir des résultats plus stables et plus nets quant à leur apport en RI. Pour ce faire, nous proposons une méthode d'acquisition d'informations morphologiques, plus souple et plus adaptée aux contraintes de la RI. Cette méthode et les résultats de son exploitation en RI sont présentés dans le chapitre suivant.

Chapitre 5

Nouvelle approche d'acquisition de variantes morphologiques utilisées pour l'extension de requêtes

Résumé : Les expérimentations précédentes ayant fait ressortir l'intérêt d'exploiter des connaissances appartenant au niveau morphologique de la langue, nous proposons au sein de ce chapitre une nouvelle façon d'aborder le problème de la variation morphologique en RI. À l'inverse de nombreux travaux existants, la technique proposée pour la reconnaissance de variantes morphologiques présente la particularité de ne nécessiter aucune ressource externe et d'être applicable à une grande variété de langues. Ces variantes servent ensuite à enrichir des requêtes. Différentes expériences sont réalisées sur plusieurs collections de documents, sur différentes langues et comparées à divers outils existants. Les résultats obtenus attestent de l'intérêt de la méthode proposée qui se révèle particulièrement bien adaptée aux spécificités de la RI.

Mots-clés : variation morphologique, extension de requêtes, apprentissage automatique par analogie.

5.1 Introduction

Au fil des différentes expériences présentées dans le cadre de cette thèse, et plus particulièrement de celles proposées pour évaluer la pertinence du couplage de connaissances multi-niveaux en RI, certaines informations linguistiques se sont révélées plus efficaces que d'autres pour permettre au SRI qui les intègrent de retrouver des documents pertinents. Qu'elles soient prises en compte de manière individuelle ou couplées avec d'autres types d'informations, ces connaissances, essentiellement de nature morphologique et prenant plus particulièrement la forme de lemmes ou de racines, contribuent en effet de manière importante à l'amélioration des systèmes. Les différents résultats obtenus attestent de leur efficacité, comme par exemple ceux présentés en section 3.4.2

qui montrent une hausse significative de la précision moyenne d'un SRI intégrant des racines par rapport à une indexation traditionnelle basée sur les termes simples.

Le traitement de la variation morphologique en RI n'est pas nouveau. Comme nous l'avons observé précédemment (*cf.* l'état de l'art sur l'apport de connaissances morphologiques en RI présenté au chapitre 2), de nombreux travaux y ont été consacrés. Plusieurs problèmes peuvent toutefois être mis au jour. Un premier est lié aux techniques proposées pour la détection des variantes morphologiques. Leur principale limite réside, pour la plupart, dans leur absence de portabilité. Elles sont en effet développées pour une langue donnée et s'appuient sur des connaissances externes (règles de réécriture, bases de suffixes, lexique...) ce qui restreint leur réutilisabilité hors de leur cadre de développement. Un second problème concerne l'exploitation même des variantes morphologiques en RI. Comme nous l'avons déjà observé, les résultats obtenus suite à leur intégration au sein de SRI sont très variables ou plus exactement dépendants de multiples paramètres comme le type de collection utilisé (longueur des requêtes, taille des collections), la langue prise en compte, ou le type d'outil employé pour leur reconnaissance. Les expériences du chapitre précédent, qui présentaient les performances d'un SRI exploitant des racines, témoignent de l'irrégularité des résultats obtenus selon les requêtes prises en compte (*cf.* l'écart-type élevé de cet index par rapport à la moyenne obtenue sur les 50 requêtes (section 4.4.3)).

Étant donné l'impact que ces informations sont susceptibles d'avoir sur les performances des systèmes, il nous semble important de disposer d'outils qui, d'une part, s'affranchissent des limites liées aux techniques d'acquisition et qui, d'autre part, soient capables de s'adapter aux spécificités des données manipulées dans une application de RI, telles que leur langue, leur taille... Notre hypothèse est que l'exploitation en RI d'informations morphologiques obtenues par de tels outils doit permettre d'obtenir des résultats plus significatifs et plus tranchés.

Pour valider cette hypothèse, nous avons envisagé, en nous appuyant sur des techniques issues de l'apprentissage artificiel, une approche simple et efficace pour la détection des variantes morphologiques, qui se fonde sur les trois hypothèses suivantes :

- la méthode d'acquisition ne doit nécessiter aucune connaissance ou données externes ;
- elle doit être complètement automatique ;
- elle doit pouvoir s'appliquer directement à diverses langues.

Les variantes acquises par le biais de notre méthode ne sont pas destinées à être normalisées pour être utilisées ensuite en RI lors de l'indexation ; la méthode d'acquisition proposée est conçue pour l'expansion de requêtes. Le but de notre approche est de détecter au sein de la collection de textes des mots en relations morphologiques avec les termes de la requête et de les ajouter à cette dernière.

Ce chapitre s'organise donc de la manière suivante : nous proposons tout d'abord (section 5.2) de positionner précisément notre approche par rapport aux travaux de l'état de l'art. Nous détaillons ensuite (section 5.3) la méthode d'acquisition envisagée et son utilisation pour l'extension de requêtes, puis présentons les différentes expérimentations réalisées sur des collections de documents et requêtes pour son évaluation (section 5.4). Nous terminons (section 5.5) par une discussion des divers résultats obtenus.

5.2 Positionnement

Notre objectif, dans cette section, est de situer précisément notre approche par rapport aux travaux qui se sont intéressés au phénomène de la variation morphologique en RI et qui ont déjà été, pour la plupart, présentés au chapitre 2 de ce mémoire. Parmi ceux-ci, peu sont compatibles avec les trois hypothèses données ici en introduction comme cadre de notre approche. En effet, la plupart des outils utilisés (*stemmer*, lemmatiseur...) reposent, comme nous l'avons vu, sur des connaissances externes, telles que des listes d'affixes, des règles de réécriture ou des dictionnaires. Ils ne peuvent par conséquent être appliqués qu'à une seule langue (*i.e.* la langue des ressources) et présentent une couverture non exhaustive. Nous proposons dans un premier temps (section 5.2.1) de présenter les quelques travaux qui s'appuient, pour la prise en compte des variantes morphologiques en RI, sur des méthodes d'acquisition qui répondent strictement à nos trois contraintes, puis décrivons ensuite quelques spécificités de notre approche par rapport à l'ensemble des travaux de l'état de l'art.

5.2.1 Travaux connexes

Les quelques travaux qui s'inscrivent dans notre cadre s'appuient, pour la prise en compte des variantes morphologiques, sur des techniques essentiellement statistiques, qui présentent l'avantage d'être indépendantes des langues et de fonctionner de manière non supervisée. Plusieurs outils de segmentation de mots (désaffixeurs) ont ainsi été développés en se basant principalement sur des critères de fréquence (Goldsmith *et al.*, 2001; Oard *et al.*, 2001). Le système AUTOMORPHOLOGY proposé par Goldsmith (2001) consiste par exemple à apprendre des listes de suffixes en s'appuyant sur la co-fréquence, dans le corpus analysé, de mots ayant une séquence commune. Néanmoins, l'application de ces techniques en RI ne semble apporter que des gains limités (Goldsmith *et al.*, 2001), ce qui est essentiellement dû à un nombre important d'erreurs engendrées (erreurs de sur- ou sous-racinisation), comme en témoignent par exemple les évaluations de Namer (2000) sur un corpus du français. D'autres approches statistiques s'appuient sur une technique de n-grammes pour la constitution de familles morphologiques (Freund et Willett, 1982; Frakes, 1992). Elles consistent plus précisément à regrouper au sein d'une même classe les mots qui partagent des sous-chaînes de caractères de longueur n, par exemple la famille de mots *juggle, juggling* et *jugglers* qui partagent le 5-gramme *juggl*. Dans certains travaux, ces classes sont affinées en utilisant du *clustering* (Frakes, 1992). Ces familles de mots peuvent ensuite être utilisées pour l'extension de requêtes. Les techniques de n-grammes s'avèrent toutefois un peu brutales, et leur utilisation en RI n'apporte en pratique pas ou peu d'améliorations (Savoy, 2002). D'une manière générale, ces méthodes essentiellement statistiques, bien que répondant à nos trois hypothèses, apparaissent peu fiables pour la détection de variantes morphologiques[1], et l'apport en

[1]La principale limite des méthodes statistiques repose sur le fait qu'elle traite le problème de la variation morphologique comme un simple regroupement de chaînes de caractères. Un tel mécanisme provoque très souvent, comme nous l'avons déjà évoqué, des erreurs (sur- et sous-racinisation).

RI des informations morphologiques obtenues par leur biais n'a pas véritablement été démontré.

5.2.2 Spécificités de l'approche proposée

Comme nous venons de le voir, la principale particularité de notre approche est de s'appuyer sur une méthode d'acquisition reposant sur les trois hypothèses présentées en introduction. Elle vise toutefois à dépasser les faiblesses des seules méthodes comparables statistiques que nous venons d'évoquer. Notre approche se distingue de celles de l'état de l'art présenté au chapitre 2 également par deux autres aspects. Le premier est directement lié à la méthode d'acquisition utilisée. La plupart des approches existantes pour la détection des variantes morphologiques s'appuient, comme nous l'avons vu, sur des techniques de désuffixation, ce qui signifie que les différentes formes d'un même mot sont reconnues comme variantes si seul leur suffixe diffère. En plus des suffixes, nous proposons, comme nous allons le voir, de prendre en compte les préfixes, ce qui permet la découverte de certaines relations morphologiques intéressantes (*e.g.* *foudre-parafoudre*, *cancéreux-anticancéreux*). L'intérêt de cette préfixation pour notre tâche de RI est évalué plus précisément en section 5.4.3. Le second point concerne la façon dont sont prises en compte les variantes morphologiques en RI. Les approches de traitement de la variation morphologique décrites dans l'état de l'art fonctionnent principalement par conflation. Dans ce cas, nous l'avons vu, les différentes variantes possibles d'un même mot sont ramenées à une forme unique qui est alors utilisée lors de l'indexation. Pour notre part, nous utilisons les variantes morphologiques uniquement pour étendre les requêtes. Bien qu'il soit difficile d'évaluer laquelle de ces deux approches — par conflation ou par expansion — est la plus efficace pour la prise en compte des variantes morphologiques en RI, l'extension de requêtes nous paraît être une technique plus souple et mieux adaptée à notre méthode (voir la discussion en section 5.5 à ce sujet). Nous rejoignons en ce sens les conclusions de Bilotti *et al.* (2004) qui, après avoir comparé les deux approches dans le cadre d'un système de question-réponse, constatent une supériorité des méthodes par expansion par rapport à une indexation basée sur les *stems* de Porter en termes de rappel, ou celles de Xu et Croft (1998) qui privilégient également l'extension de requêtes, à condition que l'approche mise en œuvre ne soit pas trop coûteuse (Xu et Croft, 1998).

Nous venons de décrire brièvement les spécificités de notre approche et son positionnement par rapport aux travaux existants. La section suivante présente plus précisément la méthode mise en œuvre pour l'acquisition de variantes morphologiques.

5.3 Acquisition de variantes morphologiques pour la RI

Afin de bien comprendre les principes de base de notre méthode d'acquisition de variantes morphologiques, nous explicitons succinctement pour commencer la façon dont elle va être concrètement utilisée. Cette méthode extrayant les variantes des documents présente la particularité, pour répondre aux trois contraintes énumérées en introduction, de s'appuyer sur une technique simple et suffisamment souple pour s'adapter aux spécificités de la RI. Par le biais de cette technique particulière (décrite ci-après), elle peut

détecter tous les couples de mots qui sont en relation morphologique (*i.e.* unis par un lien de variation morphologique). Puisque notre objectif est d'étendre les requêtes, nous l'utilisons pour repérer au sein des documents de la collection s'il existe des mots qui peuvent former un couple en relation morphologique avec l'un des termes de la requête. Si une telle paire est détectée, le mot de la collection qui appartient au couple est alors ajouté à la requête pour son enrichissement.

Nous revenons plus précisément dans un premier temps (en section 5.3.1) sur la méthode d'acquisition proposée et détaillons ensuite (en section 5.3.2) son utilisation effective au sein d'un SRI pour étendre les requêtes avec les variantes morphologiques détectées.

5.3.1 Acquisition par analogie

L'approche adoptée pour acquérir les variantes morphologiques des mots contenus dans la requête s'appuie sur une technique développée initialement à des fins terminologiques (Claveau et L'Homme, 2005b,a). Le principe de cette technique d'acquisition morphologique est relativement basique et s'appuie sur la construction d'analogies. En toute généralité, une analogie peut être représentée formellement par la proposition A : B $\doteq$ C : D, qui signifie « A est à B ce que C est à D » ; le couple A-B est donc en analogie avec le couple C-D. L'utilisation de l'analogie en morphologie, assez évidente, a déjà fait l'objet de plusieurs travaux (Hathout, 2001; Lepage, 2003). Par exemple, si l'on postule l'analogie

$$\textit{connecteur} : \textit{connecter} \doteq \textit{éditeur} : \textit{éditer}$$

et si l'on sait par ailleurs que *connecteur* et *connecter* partagent un lien morpho-sémantique, on peut alors supposer qu'il en est de même pour *éditeur* et *éditer*.

Le préalable essentiel à l'utilisation effective de l'apprentissage par analogie est la définition de la notion de similarité qui permet de statuer que deux paires de propositions — dans notre cas deux couples de mots — sont en analogie. La notion de similarité que nous utilisons, notée *Sim*, est simple mais adaptée aux nombreuses langues dans lesquelles la flexion et la dérivation sont principalement obtenues par préfixation et suffixation.

Intuitivement, *Sim* vérifie que pour passer d'un mot m_3 à un mot m_4 les mêmes opérations de préfixation et de suffixation que pour passer de m_1 à m_2 sont nécessaires. Plus formellement, notons *lcss(X,Y)* la plus longue sous-chaîne commune à deux chaînes de caractères X et Y (*e.g.* lcss(*installer*, *désinstallation*) = *install*), et X $+_{suf}$ Y (respectivement $+_{pre}$) la concaténation du suffixe (resp. préfixe) Y à X, et X $-_{suf}$ Y (respectivement $-_{pre}$) la soustraction du suffixe (resp. préfixe) Y à X. La mesure de similarité *Sim* est alors définie de la manière suivante :

Sim(m_1-m_2, m_3-m_4) = 1 si on a simultanément les quatre conditions :

$$\begin{cases} m_1 = \text{lcss}(m_1, m_2) +_{pre} \text{Pre}_1 +_{suf} \text{Suf}_1, \text{ et} \\ m_2 = \text{lcss}(m_1, m_2) +_{pre} \text{Pre}_2 +_{suf} \text{Suf}_2, \text{ et} \\ m_3 = \text{lcss}(m_3, m_4) +_{pre} \text{Pre}_1 +_{suf} \text{Suf}_1, \text{ et} \\ m_4 = \text{lcss}(m_3, m_4) +_{pre} \text{Pre}_2 +_{suf} \text{Suf}_2 \end{cases}$$

$Sim(m_1\text{-}m_2, m_3\text{-}m_4) = 0$ sinon

où Pre_i et Suf_i sont des chaînes de caractères quelconques. Si $Sim(m_1\text{-}m_2, m_3\text{-}m_4) = 1$, cela signifie que l'analogie $m_1 : m_2 \doteq m_3 : m_4$ est vérifiée ; on suppose alors que la relation morpho-sémantique entre m_1 et m_2 est la même qu'entre m_3 et m_4.

Notre processus de détection de variantes morphologiques consiste ainsi à vérifier, au moyen de la mesure *Sim*, si un couple de mots inconnus est en analogie avec un ou plusieurs exemples de couples connus. Par exemple, nous pouvons déterminer que le couple *déshydrater-réhydratation* est en analogie avec le couple-exemple *désinstaller-réinstallation*, puisque la mesure de similarité représentée comme suit :

$$\begin{cases} m_1 = \text{install}(\text{désinstaller}, \text{réinstallation}) +_{\text{pre}} \text{dés} +_{\text{suf}} \text{er}, \text{ et} \\ m_2 = \text{install}(\text{désinstaller}, \text{réinstallation}) +_{\text{pre}} \text{ré} +_{\text{suf}} \text{ation}, \text{ et} \\ m_3 = \text{hydrat}(\text{déshydrater}, \text{réhydratation}) +_{\text{pre}} \text{dés} +_{\text{suf}} \text{er}, \text{ et} \\ m_4 = \text{hydrat}(\text{déshydrater}, \text{réhydratation}) +_{\text{pre}} \text{ré} +_{\text{suf}} \text{ation} \end{cases}$$

vaut 1, ce qui s'écrit : $Sim($*désinstaller-réinstallation*, *déshydrater-réhydratation*$) = 1$.

En pratique, pour des raisons d'efficacité lors de la recherche d'analogie, plutôt que les couples-exemples, ce sont les opérations de préfixation et suffixation à l'œuvre dans la mesure de similarité *Sim* qui sont stockées. Ainsi, le couple-exemple *désinstaller-réinstallation* n'est pas stocké en tant que tel, mais on conserve la règle :

$$m_2 = m_1 -_{pre} \textit{dés} +_{pre} \textit{ré} -_{suf} \textit{er} +_{suf} \textit{ation}$$

Montrer l'analogie *déshydrater* : *réhydratation* $\doteq$ *désinstaller* : *réinstallation* revient alors simplement à tester que *déshydrater-réhydratation* vérifie la règle précédente.

Comme le soulignait Gaussier pour ses travaux (Gaussier, 1999), les opérations de suffixation et de préfixation en œuvre dans notre approche permettent de gérer en partie les variations légères de racines, pourvu qu'elles soient assez communes pour être présentes dans un de nos exemples. Si l'on a ainsi dans notre base d'exemples le couple *récupère-récupération*, le couple *agglomère-agglomération* sera facilement reconnu, malgré le changement d'accent *agglomér-/agglomèr-*, puisque l'on a bien l'analogie *récupère* : *récupération* $\doteq$ *agglomère* : *agglomération*. Bien entendu, les variations plus importantes comme celle existant dans le couple *foie-hépatique* sont hors de portée de notre approche.

Les travaux de Claveau et L'Homme (2005a) ont démontré l'efficacité de cette technique dans un contexte de construction de terminologies. Elle permet en effet de trouver des variantes morphologiques à l'aide d'exemples de mots en relation morpho-sémantique avec une très bonne couverture et une grande précision. Outre la détection de ces variantes, les auteurs montrent également qu'il est possible d'identifier avec d'excellents taux de réussite le lien sémantique précis qui les unis en associant à chaque règle[2] une ou plusieurs étiquettes de relation sémantique. Compte tenu de nos contraintes,

[2]Nous rappelons que, dans ce contexte, une règle reflète la façon dont *Sim* est calculée, c'est-à-dire la suite d'opérations nécessaires pour aller d'un terme à un autre dans un couple, *e.g.* la règle $m_2 = m_1 -_{pre} \textit{dés} +_{pre} \textit{ré} -_{suf} \textit{er} +_{suf} \textit{ation}$ pour le couple-exemple *désinstaller-réinstallation* présenté ci-dessus.

ce dernier point ne sera pas utilisé ici ; nous faisons l'hypothèse que tous les liens sémantiques (synonymie, hyperonymie, antonymie...) sont pertinents pour l'extension de requêtes.

L'application de cette technique d'acquisition de variantes morphologiques à la RI, et plus précisément pour la tâche d'extension de requêtes, nécessite quelques adaptations, sur lesquelles nous revenons à présent.

5.3.2 Utilisation en RI

Pour pouvoir être appliquée dans un contexte de RI, la méthode d'acquisition de variantes morphologiques que nous proposons doit être entièrement automatique et doit pouvoir être utilisée facilement pour étendre les requêtes de l'utilisateur. Nous détaillons donc à présent ces deux points.

5.3.2.1 Constitution automatique de couples-exemples

La technique de détection de variantes morphologiques présentée ci-dessus nécessite, pour pouvoir fonctionner, des exemples de couples de mots morphologiquement liés. Cet aspect supervisé n'est pas adapté à une utilisation en RI et à nos hypothèses exposées en introduction. Nous voulons en effet un système qui soit entièrement autonome. Pour répondre à ce problème, nous nous appuyons sur une méthode assez rustique qui permet de constituer automatiquement un ensemble de paires de mots pouvant servir de couples-exemples. Cette phase de recherche de couples-exemples se déroule de la façon suivante :

1 – choisir un document au hasard dans la collection de textes du SRI ;

2 – constituer tous les couples de mots possibles (issus du document) ;

3 – ajouter aux exemples les couples m_1-m_2 tels que $\text{lcss}(m_1,m_2) > l$;

4 – retourner en 1.

On répète ces étapes jusqu'à obtenir un ensemble de couples-exemples jugé suffisamment important. Dans les expériences présentées en section 5.4, ce sont 500 documents qui ont ainsi été analysés.

Cette phase de constitution de couples-exemples repose donc sur la même hypothèse que précédemment : la dérivation et la flexion se font principalement par des opérations de préfixation et de suffixation. En ce sens, elle est adaptée uniquement aux langues qui s'appuient sur ce type de mécanismes pour la formation des mots.

La technique utilisée pour construire nos couples-exemples ne permet pas de détecter toutes les paires de mots morphologiquement liées. Ce n'est pas problématique dans notre cas puisqu'il s'agit d'une première étape et que les couples morphologiquement liés seront retrouvés par la méthode d'acquisition par analogie. Il est en revanche primordial, pour le bon fonctionnement des analogies qui vont en être tirées, d'éviter de constituer des couples qui ne seraient pas des exemples valides. Dans notre approche simple, deux précautions ont donc été prises. D'une part, la longueur minimale de la sous-chaîne commune l est fixée à un chiffre assez grand (dans nos expériences, $l = 7$), ce qui réduit le risque de réunir deux mots ne partageant aucun lien (*e.g.* *départ-département*). D'autre

part, comme l'ont notamment observé Xu et Croft (1998), la recherche de variantes au sein d'un même document maximise les chances que les deux mots soient issus d'une même thématique et donc d'un vocabulaire cohérent.

5.3.2.2 Utilisation pour l'extension de requêtes

Après avoir constitué les couples-exemples, il convient de procéder à l'apprentissage des règles d'analogie. Ces deux étapes accomplies, il nous est maintenant possible de vérifier si un couple de mots inconnu est en analogie avec une paire connue (un de nos couples-exemples) et de déduire ainsi si les deux mots inconnus sont liés morphologiquement, *i.e.* sont en relation de dérivation ou de flexion. Dans le cadre de notre application, les mots dont on souhaite récupérer les variantes morphologiques sont ceux constituant la requête. Pour ce faire, chaque mot d'une requête est confronté à chaque mot de la collection ; si le couple ainsi formé est en relation d'analogie avec un des couples-exemples, le mot de la collection est alors considéré comme un bon candidat pour étendre la requête. En pratique, pour des questions de rapidité, les règles d'analogie sont utilisées de manière générative : des mots sont produits à partir de chaque terme de la requête en suivant les opérations de préfixation et de suffixation indiquées dans les règles et ils sont conservés uniquement s'ils apparaissent dans l'ensemble recensant tous les termes d'indexation de la collection. L'apprentissage des règles se faisant hors-ligne, seule la recherche de variantes morphologiques des termes de la requête dans l'ensemble est faite en ligne. En pratique, dans les expériences présentées ci-après, cette recherche prend quelques dixièmes de seconde[3].

Ainsi, si l'on a la requête *pollution des eaux souterraines*, la requête étendue finalement utilisée dans le SRI sur une de nos collections[4] sera par exemple : *pollution des eaux souterraines polluants dépollution anti-pollution pollutions polluées polluent eau souterraine souterrains souterrain*.

Il est important de remarquer que, lors de l'extension, seuls les mots directement liés aux termes de la requête sont ajoutés. Cette absence volontaire de transitivité doit éviter de propager des erreurs, telles que le rapprochement des termes *approvisionnement* et *vision* dans la suite *vision* → *provision* → *provisions* → *provisionner* → *approvisionner* → *approvisionnement*...

Dans les expériences présentées dans la section suivante, ce sont en moyenne 3 variantes morphologiques par mot plein de la requête qui sont ajoutées. Nous ne procédons à aucun filtrage manuel des variantes et certaines ne sont donc pas pertinentes. Il reste cependant difficile d'évaluer le nombre d'erreurs engendrées et leur impact sur les performances du SRI. De plus, une évaluation intrinsèque de ces extensions hors de leur cadre d'utilisation ne préjugerait pas de leur influence en RI. Nous reviendrons sur ce problème lors de la discussion des résultats (section 5.5).

[3]Les expérimentations ont été réalisées sur une machine de type *Pentium Centrino*-1.5 Ghz avec 512 Mo de mémoire.

[4]Il s'agit plus précisément de la collection de données INIST décrite ci-après en section 5.4.

5.4 Expériences

Cette section présente les résultats obtenus par la méthode d'extension décrite précédemment. Plusieurs expériences ont été réalisées, notamment pour évaluer la portabilité de notre technique d'acquisition et sa capacité à s'adapter aux spécificités des données manipulées en RI. Nous souhaitons en effet vérifier que les résultats obtenus sont stables quels que soient la longueur des requêtes, la taille, le type ou la langue des collections de documents utilisés. Nous décrivons dans un premier temps les résultats de notre technique d'expansion appliquée sur deux collections françaises (section 5.4.1), puis sur une collection anglaise (section 5.4.2). Nous étudions ensuite l'impact de la prise en compte des préfixes dans notre technique de détection des variantes (section 5.4.3). Nous nous intéressons (section 5.4.4) à l'influence de la longueur des requêtes sur les résultats obtenus et terminons la description de ces expériences en évaluant la portabilité de notre approche sur 6 langues différentes (section 5.4.5). Nous proposons enfin (section 5.4.6) quelques exemples de requêtes enrichies à l'aide de variantes morphologiques détectées par notre méthode.

Pour ces expérimentations, nous utilisons plusieurs collections de documents. Pour les expériences réalisées sur la langue anglaise (section 5.4.2), nous nous appuyons sur un sous-ensemble de la collection TIPSTER utilisée dans les expérimentations des deux chapitres précédents. Pour rappel, cette collection est composée d'environ 175000 documents issus du *Wall Street Journal* — couvrant des domaines liés à l'actualité — et d'un jeu de 50 requêtes. Nous nous limitons ici à l'utilisation du seul champ titre de ces requêtes, uniquement composé de quelques mots, afin de nous approcher d'un fonctionnement « grand public » (*i.e.* requêtes courtes).

Les autres expérimentations ont été menées pour le français à partir de la collection de données INIST composée de 30 requêtes et de 163 308 documents, résumés d'articles relevant de différentes disciplines scientifiques, et de la collection ELRA composée de 30 requêtes et 3511 documents issus de questions/réponses de la commission européenne. Pour les expériences évaluant la portabilité de notre approche sur plusieurs langues à la fois (section 5.4.5), nous utilisons également la collection ELRA qui présente la particularité d'être disponible en français, anglais, allemand, portugais, espagnol et italien. Dans ces deux collections (ELRA et INIST), les requêtes comportent plusieurs champs (titre, question, informations complémentaires, concepts associés). Comme pour l'anglais, nous faisons le choix d'utiliser des requêtes contenant peu de mots ; les questions effectivement utilisées sont composées uniquement du contenu du champ titre, sauf en section 5.4.4 où nous étudions l'influence de la taille des requêtes sur notre technique d'extension.

Le système de recherche d'information que nous utilisons pour ces expériences est LEMUR, déjà employé au chapitre 3, paramétré de manière à adopter le fonctionnement du système OKAPI (Robertson *et al.*, 1998).

5.4.1 Résultats sur le français

Cette première expérience est réalisée sur le français à partir des collections INIST et ELRA, en utilisant pour toutes les deux des requêtes courtes (champ sujet). Pour évaluer l'apport de l'extension de requêtes à l'aide des variantes morphologiques acquises avec notre méthode, nous comparons les résultats obtenus avec et sans cette extension, mesurés en termes de précision et rappel à différents seuils, de précision moyenne interpolée sur 11 points (IAP), de R-précision et de précision moyenne non interpolée (MAP) (ces mesures sont détaillées en section 1.4.2).

À titre de comparaison, nous présentons également les résultats obtenus par deux systèmes standards manipulant des informations morphologiques, usuellement utilisés en RI pour leur robustesse et leur disponibilité : un *stemmer* du français (développé par Savoy (1999), s'appuyant sur un ensemble de règles fixes de désuffixation), et un lemmatiseur du français (l'étiqueteur TREETAGGER) déjà utilisé lors des expérimentations précédentes. Ces deux outils fonctionnent par conflation, c'est-à-dire qu'ils ramènent les mots à leur forme canonique (*stem* ou lemme) qui est utilisée lors de l'indexation des documents et requêtes.

Les tableaux 5.1 et 5.2 récapitulent les résultats respectivement obtenus pour la collection INIST et ELRA ; les chiffres jugés statistiquement non significatifs par un *paired t-test* (Hull, 1993) (avec une valeur $p < 0.05$) apparaissent en petites italiques. La taille moyenne des requêtes ($|Q|$) est également indiquée, calculée en nombre de mots (y compris les mots-vides).

Il ressort de ces deux tableaux que notre approche obtient de très bons résultats pour chacune des mesures adoptées, tous statistiquement significatifs à deux exceptions près dans la collection ELRA. Pour la plupart des mesures, l'extension de requêtes est notamment plus performante que le *stemming* ou la lemmatisation, mais aussi plus stable comme l'atteste certains résultats jugés non statistiquement significatifs de ces deux techniques. Les résultats semblent largement dépendants de la collection utilisée. À ce titre, notre approche paraît néanmoins assez robuste, contrairement à la lemmatisation qui n'apporte aucune amélioration pour la collection ELRA, ou au *stemming* dont l'effet n'est sensible (et restreint) qu'après les 500 premiers documents. Il est également intéressant de noter que l'amélioration des résultats est également distribuée sur tous les seuils de mesures (de 10 à 5 000 documents), quelle que soit la collection utilisée. Cela signifie que l'amélioration n'est pas due à une simple réorganisation des documents en tête de liste mais plutôt à la découverte de documents pertinents qui n'auraient pas été ramenés par le SRI sans les extensions de requêtes.

5.4.2 Résultats sur l'anglais

L'objectif de cette deuxième expérience est de déterminer si les bons résultats obtenus par notre approche sur le français peuvent également être observés sur l'anglais. Pour cela, nous réitérons les expérimentations précédentes, en nous appuyant cette fois sur la collection anglaise TIPSTER. Le tableau 5.3 présente les résultats obtenus sur cette

	Sans extension	Avec extension (amélioration %)	*Stemming* (amélioration %)	Lemmatisation (amélioration %)
|Q|	5.3	16.03	5.2	5.17
MAP	14.85	18.45 (+24.29%)	*17.31 (+16.63%)*	17.82 (+20.07%)
IAP	16.89	19.93 (+17.97%)	*18.85 (+11.57%)*	19.72 (+16.73%)
R-Prec	17.99	21.63 (+20.24%)	*19.88 (+10.53%)*	*19.71 (+9.56%)*
P(10)	34.33	38.67 (+12.62%)	*36.67 (+6.80%)*	39.67 (+15.53%)
P(20)	27.83	31.83 (+14.37%)	*29.00 (+4.19%)*	31.6 (+13.77%)
P(50)	18.33	21.27 (+16.00%)	*20.13 (+9.82%)*	*20.87 (+13.82)*
P(100)	12.23	14.80 (+20.98%)	15.23 (+24.52%)	14.97 (+22.34%)
P(200)	8.02	9.73 (+21.41%)	9.77 (+21.82%)	9.5 (+18.50%)
P(500)	3.88	4.80 (+23.71%)	4.55 (+17.18%)	4.47 (+15.29%)
P(1 000)	2.21	2.68 (+21.30%)	2.53 (+14.80%)	2.48 (+12.39%)
P(2 000)	1.29	1.50 (+16.45%)	1.40 (+8.68%)	1.39 (+8.29%)
P(5 000)	0.56	0.67 (+20.38%)	0.63 (+13.47%)	0.64 (+15.14%)
R(10)	8.00	8.99 (+12.36%)	*8.45 (+5.64%)*	*9.04 (+13.02%)*
R(20)	12.33	14.50 (+17.59%)	*12.81 (+3.90%)*	*13.62 (+10.48%)*
R(50)	19.65	24.07 (+22.47%)	*20.78 (+5.74%)*	*21.56 (+9.71%)*
R(100)	26.85	32.87 (+22.41%)	31.32 (+16.64%)	31.58 (+17.59%)
R(200)	34.38	42.70 (+24.21%)	41.05 (+19.43%)	40.92 (+19.04%)
R(500)	43.09	53.83 (+24.92%)	49.31 (+14.43%)	49.35 (+14.54%)
R(1 000)	48.43	59.45 (+22.74%)	55.27 (+14.12%)	55.03 (+13.62%)
R(2 000)	55.35	65.85 (+18.99%)	61.08 (+10.36%)	61.20 (+10.57%)
R(5 000)	59.32	72.20 (+21.71%)	67.22 (+13.31%)	68.20 (+14.96%)

TAB. 5.1 – Performances de l'extension de requêtes sur la collection INIST

collection par notre méthode comparés à ceux d'une recherche classique sans extension. L'amélioration relative (en pourcentage) est indiquée entre parenthèses.

Nous pouvons constater à travers ce tableau que les résultats obtenus sont également tous positifs. L'apport de notre méthode d'extension de requêtes sur l'anglais est conséquent puisqu'une amélioration des performances, comprise entre 7 et 17% selon les mesures, est observée. Comme précédemment, nous pouvons remarquer que le gain obtenu concerne tous les seuils de mesures de précision et de rappel ; l'extension de requêtes a donc permis de détecter de nouveaux documents pertinents.

Ces observations mettent en évidence la robustesse de notre méthode et sa capacité à s'adapter aux spécificités de la langue prise en compte. Plusieurs expérimentations sont proposées ci-après (section 5.4.5) pour évaluer la portabilité de notre approche sur d'autres langues. Au regard de ces premiers résultats, nous pouvons déjà affirmé que l'apport de la prise en compte de variantes morphologiques en RI par le biais de notre méthode ne semble pas lié à la complexité morphologique de la langue considérée,

	Sans extension	Avec extension (amélioration %)	*Stemming* (amélioration %)	Lemmatisation (amélioration %)
|Q|	2.83	9.23	2.8	2.83
MAP	42.26	47.29 (+11.89%)	*42.63 (+0.87%)*	*41.32 (-2.24%)*
IAP	43.14	47.88 (+10.99%)	*44.05 (+2.11%)*	*42.58 (-1.3%)*
R-Prec	44.22	48.39 (+9.43%)	*45.1 (+2.1%)*	*42.89 (-3.01%)*
P(10)	56.67	*58.67 (+3.53%)*	*54.67 (-3.53%)*	*56.33 (-0.59%)*
P(20)	48.67	52 (+6.85%)	*48.17 (-1.03%)*	*49 (+0.68%)*
P(50)	30.33	34.53 (+13.85%)	*30.4 (+0.22%)*	*29.13 (-3.96%)*
P(100)	19.7	22.9 (+16.24%)	*20.1 (+2.38%)*	*18.1 (-8.12%)*
P(200)	11.42	12.87 (+12.7%)	*11.58 (+2.81%)*	*10.65 (-6.72%)*
P(500)	4.84	5.68 (+17.19%)	4.95 (+5.99%)	*4.64 (-4.26%)*
P(1 000)	2.45	2.89 (+18.26%)	2.52 (+6.64%)	*2.24 (-4.22%)*
P(2 000)	1.22	1.45 (+18.66%)	1.27 (+7.49%)	*1.17 (-4.22%)*
R(10)	22.36	*23.57 (+5.43%)*	*22.27 (-0.37%)*	*22.44 (+0.37%)*
R(20)	35.55	38.16 (+7.36%)	*37.73 (+6.14%)*	*37.39 (+5.19%)*
R(50)	49.30	54.64 (+10.82%)	*51.41 (+4.28%)*	*49.98 (+1.39%)*
R(100)	60.04	66.10 (+10.11%)	*62.12 (+3.64%)*	*58.27 (-2.95%)*
R(200)	68.52	73.26 (+6.92%)	*70.84 (+4.08%)*	*66.96 (-2.27%)*
R(500)	72.50	81.69 (+12.68%)	76.37 (+7.29%)	*72.53 (+0.05%)*
R(1 000)	73.42	83.11 (+13.2%)	77.56 (+7.57%)	*73.83 (+0.55%)*
R(2 000)	73.42	83.48 (+13.7%)	78.19 (+8.43%)	*73.83 (+0.55%)*

TAB. 5.2 – Performances de l'extension de requêtes sur la collection ELRA

contrairement à ce qui est parfois constaté dans les travaux du domaine (*cf.* section 2.2.2).

Ces résultats confirment également les conclusions du chapitre 3 sur l'intérêt de recourir de manière plus générale à un traitement morphologique en RI. En effet, quelle que soit la façon d'exploiter les connaissances morphologiques au sein des systèmes — par le biais d'une approche par conflation lors de l'indexation comme dans le chapitre 3[5] ou par expansion lors de la recherche comme ici — l'apport de leur prise en compte en RI est indéniable.

5.4.3 Influence de la prise en compte des préfixes

Contrairement à la plupart des approches traitant le phénomène de la variation morphologique en RI, notre méthode prend en compte l'opération de préfixation. Bien que cette opération semble intéressante dans certains cas (comme dans l'exemple de *foudre-parafoudre*), elle peut également introduire du bruit (couples de mots non liés

[5] *Cf.* les résultats obtenus par le SRI exploitant un index de racines ou de lemmes dans les expériences présentées en section 3.4.2.

	Sans extension	Avec extension (amélioration %)
|Q|	6.5	16.6
MAP	23.10	26.60 (+14.91%)
IAP	25.02	28.68 (+14.60%)
R-Prec	27.52	31.09 (+12.98%)
P(10)	39.60	46 (+16.16%)
P(20)	36.10	41.20 (+14.12%)
P(50)	28.28	32.40 (+14.56%)
P(100)	21.44	24.34 (+13.52%)
P(200)	14.52	17.06 (+17.49%)
P(500)	7.94	9.09 (+14.39%)
P(1 000)	4.66	5.26 (+12.81%)
P(2 000)	2.59	2.96 (+14.06%)
R(10)	10.20	10.98 (+7.65%)
R(20)	16.10	18.07 (+12.22%)
R(50)	29.68	32.09 (+8.12%)
R(100)	39.48	43.66 (+10.59%)
R(200)	49.28	54.99 (+11.57%)
R(500)	61.11	67.16 (+9.91%)
R(1 000)	68.68	74.64 (+8.67%)
R(2 000)	74.41	81.01 (+9.18%)

TAB. 5.3 – Performances de l'extension de requêtes sur la collection TIPSTER

sémantiquement ou faiblement liés) néfaste pour l'expansion. En effet, l'ajout d'un préfixe, en français, signe d'une dérivation, induit souvent une modification de sens (*e.g.* *pondre-répondre*) plus importante que pour la suffixation (très fréquemment signe d'une flexion). C'est pourquoi, l'opération de préfixation est le plus souvent ignorée dans les traitements morphologiques utilisés en RI (Savoy, 2002). Pour vérifier l'intérêt de la préfixation dans notre cadre, nous répétons donc les deux expériences précédentes en évaluant d'une part la méthode d'extension prenant en compte les préfixes et suffixes et, d'autre part, la même méthode limitée aux extensions trouvées par changement, ajout ou suppression de suffixes. La différence observée entre les résultats des deux méthodes nous permet ainsi d'évaluer l'intérêt de la prise en compte des préfixes. Le tableau 5.4 présente les résultats obtenus sur les collections INIST et ELRA en français. Comme précédemment, l'amélioration relative est calculée à partir des résultats sans aucune extension.

Ces expériences montrent tout d'abord que la prise en compte de la préfixation n'apporte que quelques extensions : un peu plus d'un terme par requête en moyenne. Cependant, ces extensions apportent un gain léger mais constant de performances à tous les seuils de mesures pour les deux collections. L'opération de préfixation semble donc

	INIST		ELRA	
	Préfixes et suffixes	Suffixes seulement	Préfixes et suffixes	Suffixes seulement
|Q|	16.03	14.77	9.07	7.93
MAP	18.45 (+24.29%)	17.87 (+20.34%)	47.47 (+12.30%)	47.4 (+12.14%)
IAP	19.93 (+17.97%)	19.59 (+16.06%)	48.00 (+11.22%)	47.99 (+11.21%)
R-Prec	21.63 (+20.24%)	20.79 (+15.58%)	48.54 (+9.76%)	48.74 (+10.21%)
P(10)	38.67 (+12.62%)	38.67 (+12.62%)	*58.33 (+2.94%)*	*58.33 (+2.94%)*
P(20)	31.83 (+14.37%)	16.67 (+13.77%)	51.67 (+6.16%)	51.67 (+6.16%)
P(50)	21.27 (+16.00%)	20.93 (+14.23%)	34.6 (+14.07%)	34.67 (+14.29%)
P(100)	14.80 (+20.98%)	14.47 (+18.26%)	23 (+16.75%)	23 (+16.75%)
P(200)	9.73 (+21.41%)	9.57 (+19.33%)	12.87 (+12.7%)	12.88 (+12.85%)
P(500)	4.80 (+23.71%)	4.69 (+20.58%)	5.68 (+17.19%)	5.66 (+16.78%)
P(1 000)	2.68 (+21.30%)	2.6 (+17.83%)	2.89 (+18.12%)	2.88 (+17.71%)
P(2 000)	1.50 (+16.45%)	1.46 (+13.73%)	1.45 (+18.53%)	1.44 (+17.98%)
P(5 000)	0.67 (+20.38%)	0.66 (+17.16%)	0.58 (+18.53%)	0.58 (+17.98%)
R(10)	8.99 (+12.36%)	8.76 (+9.45%)	*23.49 (+5.08%)*	*23.43 (+4.81%)*
R(20)	14.50 (+17.59%)	14.24 (+15.54%)	37.93 (+6.69%)	38.02 (+6.97%)
R(50)	24.07 (+22.47%)	22.88 (+16.46%)	54.84 (+11.24%)	54.92 (+11.4%)
R(100)	32.87 (+22.41%)	31.99 (+19.12%)	66.39 (+10.59%)	66.35 (+10.52%)
R(200)	42.70 (+24.21%)	41.95 (+22.02%)	73.39 (+7.12%)	73.42 (+7.15%)
R(500)	53.83 (+24.92%)	52.28 (21.33%)	81.48 (+12.40%)	81.24 (+12.06%)
R(1 000)	59.45 (+22.74%)	57.84 (+19.43%)	82.87 (+12.87%)	82.63 (+12.55%)
R(2 000)	65.85 (+18.99%)	64.10 (+15.81%)	83.24 (+13.38%)	82.91 (+12.93%)
R(5 000)	72.20 (+21.71%)	70.17 (+18.29%)	83.24 (+13.38%)	82.91 (+12.93%)

TAB. 5.4 – Performances de l'extension de requêtes avec et sans préfixation

être utile à notre système d'extension même si la plus grande partie des améliorations de performances est due à des variantes issues de changement, ajout ou suppression de suffixes.

5.4.4 Influence de la taille des requêtes

Les résultats mitigés obtenus par beaucoup d'outils intégrant la variation morphologique en RI sont souvent liés, nous l'avons vu, à la taille des requêtes prises en compte. Pour mesurer son influence sur notre approche, nous répétons l'expérience précédente en utilisant cette fois les autres champs des requêtes INIST pour composer des requêtes de plus en plus longues. Nous nous appuyons plus particulièrement sur les champs *concept* de ces requêtes où chaque concept est représenté sous la forme d'un mot-clé proche du contenu sémantique de la requête. Ces mots-clés sont donc ajoutés un par un à la requête, composée initialement du seul champ sujet. À titre de comparaison, cette ex-

périence est également réalisée sur un SRI basé sur un traitement de *stemming* et de lemmatisation. La figure 5.1 présente les résultats obtenus selon la taille, mesurée en nombre de mots avant toute extension, des requêtes ainsi constituées. La performance du SRI est mesurée par la précision moyenne non interpolée.

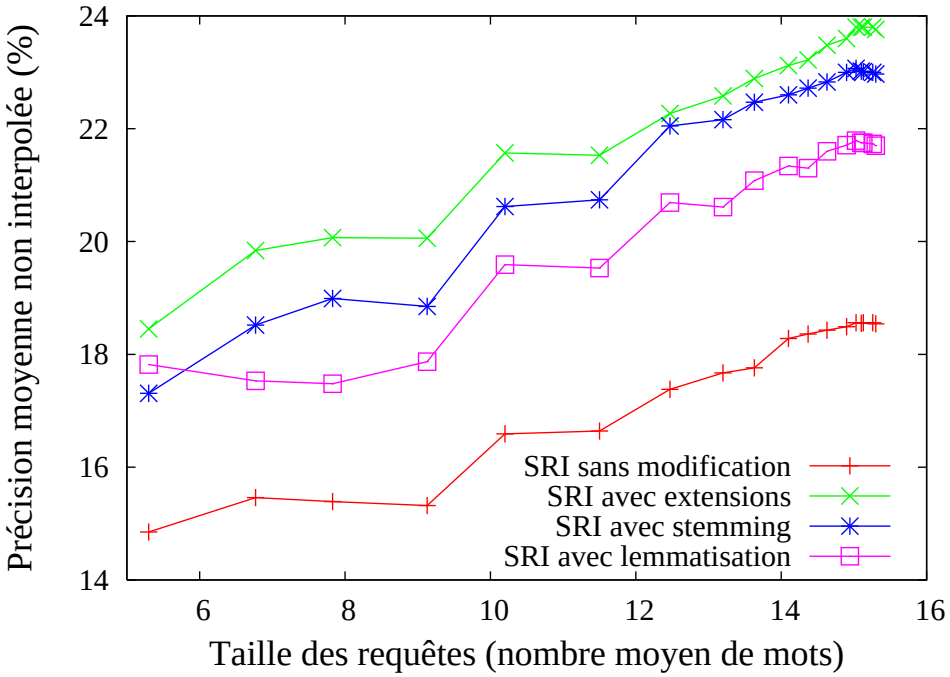

FIG. 5.1 – Évolution de la précision selon la taille de la requête

D'une manière générale, ces résultats mettent en évidence l'intérêt de gérer les variantes morphologiques en RI, et ce, quelle que soit la taille de la requête et la technique adoptée — par *stemming*, lemmatisation ou avec notre méthode d'extension. Parmi les trois méthodes comparées, notre approche d'extensions morphologiques apparaît la plus performante, devant le *stemming* et la lemmatisation. Il est également intéressant de constater que, contrairement aux résultats constatés dans certains travaux (*cf.* section 2.2.2), la prise en compte de la morphologie apporte un gain de performance constant, même pour les requêtes longues. Ces observations confirment l'idée que les connaissances d'ordre morphologique sont des informations fiables sur lesquelles on peut s'appuyer pour améliorer les performances des SRI.

5.4.5 Évaluation de la portabilité

Notre approche, n'utilisant aucune connaissance externe, se veut par conséquent portable ; elle doit donc être directement utilisable pour n'importe quelle langue dont la morphologie se fait principalement par préfixation et suffixation. Pour vérifier la portée de cette assertion, nous présentons ci-dessous dans le tableau 5.5 les résultats obtenus

sur la collection ELRA pour l'allemand, l'anglais, l'espagnol, le français, l'italien et le portugais. Pour chacune de ces langues, nous indiquons la variation (en pourcentage) par rapport à la même recherche utilisant les requêtes sans extension.

	Allemand	Anglais	Espagnol	Français	Italien	Portugais
MAP	+16.25%	+17.52%	+10.03%	+11.89%	+10.45%	+9.69%
IAP	+15.93%	+16.66%	+8.70%	+10.99%	+9.79%	+9.25%
R-Prec	*+3.03%*	+10.23%	+7.97%	+9.43%	+10.23%	*+6.20%*
P(10)	*+10.68%*	+7.03%	*0%*	*+3.53%*	*+2.54%*	*0%*
P(20)	*+8.33%*	*+3.62%*	+7.41%	+6.85%	+11.15%	*+4.38%*
P(50)	*+6.69%*	+8.23%	+13.40%	+13.85%	+13.48%	+8.31%
P(100)	+9.54%	+14.31%	+16.76%	+16.24%	+18.98%	+14.24%
P(200)	+12.61%	+19.63%	+14.15%	+12.70%	+18.70%	+19.27%
P(500)	+13.18%	+20.49%	+18.13%	+17.19%	+18.94%	+23.35%
P(1 000)	+12.97%	+21.60%	+20.32%	+18.26%	+22.13%	+24.64%
P(2 000)	+12.33%	+19.94%	+20.70%	+18.66%	+22.71%	+24.85%
R(10)	*+6.82%*	+2.90%	*+1.88%*	*+5.43%*	*-0.67%*	*-0.47%*
R(20)	*+5.95%*	*+3.27%*	+7.40%	+7.36%	+7.82%	*+7.55%*
R(50)	*+11.12%*	+8.48%	+7.72%	+10.82%	+7.37%	*+6.21%*
R(100)	+11.87%	+13.23%	+10.14%	+10.11%	+8.93%	+9.39%
R(200)	+18.23%	+19.43%	+7.40%	+6.92%	+9.77%	+14.18%
R(500)	+16.45%	+21.68%	+14.49%	+12.69%	+14.31%	+17.71%
R(1 000)	+18.15%	+20.93%	+17.38%	+13.20%	+18.35%	+19.23%
R(2 000)	+16.34%	+17.77%	+17.54%	+13.70%	+18.97	+19.26%

TAB. 5.5 – Performances de l'extension de requêtes sur différentes langues

Les résultats sont tous très positifs puisque l'extension de requêtes apporte un gain de performance de 10 à 20% selon les langues et les mesures. Comme pour le français dans les expériences précédentes, l'amélioration se fait en particulier sentir à tous les seuils de mesures de précision et de rappel ; cependant, pour les seuils bas (10 à 50 documents), les résultats sont très variables d'une requête à une autre, ce que traduit le fait qu'ils ne soient pas jugés statistiquement significatifs.

Nous observons également que l'anglais, réputée de morphologie pauvre, bénéficie plus de l'extension de requêtes par variantes morphologiques que des langues à morphologie plus riche (espagnol, italien...). Autre résultat surprenant, l'allemand est la langue bénéficiant le plus de notre technique d'extension, sans doute par le fait que notre approche capture des cas d'agglutination fréquents dans cette langue, comme par exemple le couple *Menschenrechte-Menschenrechtsorganisation*.

5.4.6 Quelques exemples de requêtes étendues

À titre d'exemple, nous proposons ici d'analyser quelques requêtes enrichies à l'aide de variantes acquises par notre technique. Ces requêtes sont issues de la collection française INIST.

requête initiale *politique énergétique et protection de l'environnement*
requête étendue *politique politisation politisée politiques énergétique énergétiques énergies énergie protection protecteur protecteurs protections protectrice protectrices environnement environnemental environnementale environnementales environnementaux environnements environnant*

Pour cette requête, nous constatons que les variantes ajoutées sont globalement pertinentes. Le rapprochement des formes *énergétique* et *énergie* est effectué par le biais de la règle suivante :

$$m_2 = m_1 -_{suf} \textit{étique} +_{suf} \textit{ie}$$

où m_1 représente ici *énergétique* et m_2 *énergie*. Notre technique est cependant passée à côté de variantes intéressantes, ne permettant pas par exemple de rapprocher le terme *protection* contenu dans la requête de sa variante *protéger* présente dans la collection de documents.

requête initiale *la mobilité et l'insertion professionnelle et sociale des femmes et des jeunes*
requête étendue *mobilité mobilités hypermobilité insertion inserts ré-insertion insertions réinsertion professionnelle professions professionnel professionnelles interprofessionnelle sociale sociaux psychosociale psycho-sociale sociales socialisé social socialement socialisation pro-sociales prosociales femmes femme jeunes jeune*

Dans cet exemple, la détection de variantes comme *réinsertion*, *ré-insertion* ou *psychosocial* est particulièrement pertinente pour la découverte de nouveaux documents, illustrant ainsi l'intérêt de prendre en compte les préfixes pour la détection de variantes morphologiques.

Certaines requêtes de notre collection contiennent un vocabulaire plus spécialisé, appartenant notamment au domaine médical, comme par exemple les deux requêtes suivantes :

requête initiale *diarrhée chez le nouveau-né et l'enfant*
requête étendue *diarrhée antidiarrhéique antidiarrhéiques diarrhéique diarrhéiques diarrhérique diarrhéogènes diarrhéigènes diarrhées nouveau-né nouveau-nées nouveau-née nouveau-nés enfant enfants enfante enfance*

requête initiale *ulcères gastroduodénaux*
requête étendue *ulcères ulcère ulcèreux ulcération ulcèrations gastroduodénaux gastroduodénale gastro-duodénale gastroduodénal gastro-duodénal gastro-duodénaux*

gastroduodénales gastro-duodénales

Nous pouvons constater à travers ces deux requêtes que notre technique d'acquisition de variantes est particulièrement intéressante sur un vocabulaire spécialisé. Les termes de la requête possèdent au sein des documents de la collection un grand nombre de variantes et les règles acquises permettent d'établir des rapprochements intéressants comme le montre par exemple le lien entre les formes *gastroduodénaux* et *gastro-duodénal* ou entre *diarrhée* et *antidiarrhéique*, obtenu par le biais de la règle :

$$m_2 = m_1 +_{pre} \textit{anti} -_{suf} \textit{e} +_{suf} \textit{ique}$$

où m_1 et m_2 représentent respectivement les formes *diarrhée* et *antidiarrhéique*.

Notre approche n'est cependant pas exempte d'erreurs, puisque, comme le montrent les deux exemples suivants, des termes non pertinents sont parfois utilisés pour l'extension.

requête initiale *sida en Afrique*
requête étendue *sida présida afrique*

Pour cette requête, notre méthode n'a détectée aucune variante valide à ajouter aux termes initiaux, allant même jusqu'à introduire du bruit par le biais de la forme *présida*.

requête initiale *marché du travail*
requête étendue *marché marchés hypermarché supermarché travail non-travail travaille télétravail*

Dans cet exemple, les erreurs observées concernent essentiellement l'ajout de préfixes qui, dans ce contexte précis, ne sont pas pertinents. Seule l'utilisation d'un traitement de désambiguïsation pourrait, en opérant une distinction entre les différents sens du terme *marché* par exemple, éviter ce type de problème. D'autre part, les expériences précédentes (section 5.4.3) ont montré l'intérêt global des opérations de préfixation. Ce type d'erreur est donc minoritaire au sein des résultats observés.

5.5 Discussions des résultats

Les expériences décrites dans ce chapitre ont permis de mettre au jour des observations importantes allant à l'encontre de conclusions parfois avancées dans les travaux traitant de la variation morphologique en RI. D'une part, les résultats obtenus ne semblent pas dépendre de la richesse morphologique de la langue prise en compte. En effet, contrairement à d'autres études (Arampatzis *et al.*, 2000), les traitements morphologiques s'avèrent efficaces pour améliorer les performances des SRI que ce soit pour des langues à morphologie « pauvre » (*e.g.* l'anglais) ou pour des langues morphologiquement plus complexes (*e.g.* l'espagnol, l'italien, le portugais...). D'autre part, la taille des

requêtes se semble pas non plus influencer dans un sens ou l'autre les résultats puisque les améliorations observées sont constantes et de même ordre de grandeur pour des requêtes allant de 5 à 15 mots. Puisque la prise en compte de la variation morphologique ne semble pas tributaire de ces deux facteurs, on peut considérer que ce traitement est fiable et efficace pour améliorer les performances des SRI.

Si les résultats sont tous positifs et constants au sein d'une même collection, on constate néanmoins une variation importante de performances entre les collections pour une même langue. Il semble donc que ce type d'approche prenant en compte la variation morphologique soit en fait sensible à la collection plus qu'à la langue. Il faut souligner à ce titre quelques particularités de la collection multilingue ELRA qui peuvent expliquer les résultats obtenus pour le français, mais aussi pour les autres langues. Tout d'abord, la collection comporte peu de documents (environ 3500 documents contre 163 308 pour la collection INIST), les requêtes sont beaucoup plus courtes — il y a donc moins de mots pleins auxquels il est possible d'ajouter des variantes —, et les types de documents qu'elle contient (débats parlementaires contre résumés d'articles scientifiques) doivent certainement avoir un impact sur la productivité morphologique.

Concernant plus spécifiquement la méthode proposée pour enrichir les termes de la requête à l'aide de leurs variantes morphologiques, plusieurs erreurs ont pu être constatées. Elle ne permet pas tout d'abord de retrouver certains termes pourtant morphologiquement liés. Ce type d'erreur n'a pas d'incidence sur les résultats obtenus si ce n'est que les améliorations constatées pourraient encore être meilleures. Il lui arrive en revanche de détecter des termes non pertinents ; le bruit introduit a alors un impact plus néfaste sur les performances des systèmes. Dans ce type d'erreur, il est nécessaire de distinguer plusieurs cas. Tout d'abord, certains mots trouvés n'entretiennent pas de liens sémantiques avec le terme original, le lien morphologique étant fortuit (*pondre-répondre*) ou pertinent d'un point de vue diachronique mais plus en œuvre dans la langue actuelle, comme par exemple *composition-exposition* ou *ordination-ordinateur*. Ensuite, certains termes polysémiques provoquent des erreurs difficiles à éviter. Ainsi, les deux mots *production* et *reproduction* trouvés comme étant liés le sont dans les phrases *production des résultats* et *reproduction des résultats*, mais pas dans *la reproduction chez les poissons*. Ces erreurs mettent en évidence l'intérêt qu'il y aurait à utiliser des outils de désambiguïsation. Il est nécessaire de rappeler que l'absence volontaire de transitivité dans nos extensions, c'est-à-dire le fait que l'on n'ajoute aux mots des requêtes que des extensions qui leurs sont directement liées (*cf.* l'exemple de *vision* et *provision*) permet de limiter ces erreurs. À ce titre, une approche par expansion apparaît comme un cadre plus souple que les techniques fonctionnant par conflation dans lesquelles *production* et *reproduction* ou *départ* et *département* et toutes leurs variantes seraient ramenés à une forme unique.

Enfin, nous avons pu remarqué que dans certains cas, les deux mots sont bien reliés sémantiquement et semblent bien correspondre au sens utilisé dans la requête, et pourtant l'ajout du terme lié dégrade les performances du SRI. Par exemple, dans la collection INIST, l'extension de la requête *impact sur l'environnement des moteurs diesel* avec le terme *moteur* fait chuter la qualité des résultats proposés. Dans ce cas, l'usage fait que, lorsque l'on emploie le terme *moteur* dans ce type de contexte, il semble que ce dernier soit la plupart du temps utilisé au pluriel.

5.6 Conclusion

Nous avons proposé une technique simple permettant de détecter automatiquement des variantes morphologiques au sein des textes en s'appuyant sur la construction d'analogie. Cette technique, qui repose sur l'apprentissage de schémas de préfixation et de suffixation, a été utilisée en RI pour étendre les requêtes à l'aide des variantes morphologiques des termes qu'elles contiennent. Les résultats obtenus par notre approche non supervisée sont très satisfaisants et se comparent avantageusement aux outils supervisés testés (*stemmer* à base de règles et lemmatiseur). Les améliorations constatées sont effet assez importantes (par exemple entre 12% et 24% pour les deux collections de textes en français) et constantes. En ce sens, notre approche diffère des travaux existants traitant la variation morphologique qui obtiennent, nous l'avons vu, des résultats généralement mitigés.

Nous avons également étudié l'intérêt de prendre en compte les préfixes dans la détection des variantes morphologiques, ces derniers étant la plupart du temps ignorés par les traitements morphologiques traditionnellement utilisés en RI. Bien que l'opération de préfixation n'apporte que peu de variantes, ces dernières apparaissent pertinentes et apportent un gain léger de performances. De plus, nous avons montré que l'aspect non supervisé de notre technique d'extension permettait de l'appliquer à diverses langues. Là encore, les expérimentations menées pour l'allemand, l'anglais, l'espagnol, le français, l'italien et le portugais attestent du bien-fondé de notre approche, avec des gains de performances importants et assez homogènes d'une langue à l'autre.

Du point de vue de la RI, nos expérimentations mettent en évidence plusieurs points importants. Elles confirment tout d'abord l'idée que les informations d'ordre morphologique contribuent de manière significative à l'amélioration des performances des systèmes, et renforcent les résultats observés lors des expérimentations décrites au chapitre 3. Dans nos expériences, ces connaissances s'avèrent fiables et non tributaires des facteurs généralement évoqués comme la taille des requêtes, la longueur des documents, la langue... L'approche d'extraction proposée a également mis en valeur l'intérêt de concevoir pour une application de RI des outils qui s'appuient sur des techniques simples et capables de s'adapter aux spécificités de cette application. Les méthodes qui exploitent essentiellement des connaissances linguistiques acquises automatiquement à partir des collections sont, nous l'avons montré, plus portables et plus souples, s'avèrent par conséquent plus efficaces en RI. Ce constat laisse penser que les outils et techniques du TAL traditionnellement utilisés pour chercher à améliorer les performances des SRI ne sont peut-être pas véritablement adaptés au domaine de la RI, car généralement conçus de manière *ad hoc*, *i.e.* indépendamment de l'application visée (la RI). Cette remarque pourrait en partie expliquer pourquoi les résultats généralement obtenus lors des expériences couplant le TAL à la RI ne sont pas plus significatifs.

Conclusion

Synthèse des travaux

Compte tenu du fossé important existant entre d'une part l'intérêt théorique qu'il y a à exploiter en RI des informations linguistiques issues de techniques du TAL pour améliorer les mécanismes de recherche traditionnels et, d'autre part, les résultats peu tranchés et plutôt contradictoires observés lors des nombreuses tentatives de couplage TAL-RI déjà réalisées quant à leur apport réel, notre objectif, présenté en introduction générale, était de comprendre pourquoi les informations issues du TAL ne donnaient pas les résultats escomptés et surtout comment procéder pour tenter de faire mieux en regardant le couplage TAL et RI sous des angles nouveaux.

Pour cela, nous avons dans un premier temps effectué une étude approfondie des multiples travaux existants ayant tenté d'insérer des connaissances linguistiques dans des SRI. Nous avons proposé une synthèse détaillée de ce domaine particulièrement vaste et hétérogène, en essayant de mettre en évidence les différentes raisons des succès ou échecs rencontrés. En nous appuyant sur cette étude et sur les constats intéressants qui ont pu y être faits, nous avons choisi d'aborder le couplage TAL-RI sous les trois nouveaux angles suivants :

- étudier l'intérêt en RI de coupler des informations linguistiques multi-niveaux. Ceci est motivé par deux raisons : les travaux existants se limitent, pour la plupart, à l'exploitation d'une seule information de niveau morphologique, syntaxique ou sémantique ; la richesse de la langue n'est exploitée que partiellement ;
- trouver le moyen de les coupler au mieux et automatiquement au sein d'un SRI ;
- trouver le moyen d'obtenir des informations linguistiques de qualité en nous appuyant sur des méthodes d'extraction qui s'adaptent aux caractéristiques et contraintes de l'application visée de RI.

Concernant le premier point, nous avons proposé une plate-forme permettant l'intégration en parallèle de multiples index linguistiques. Nous avons exploré 11 des index linguistiques les plus standards (lemmes, racines, termes étiquetés grammaticalement, groupes nominaux, bigrammes, trigrammes, termes complexes, noms propres, termes étiquetés sémantiquement, synonymes et mots reliés morpho-sémantiquement), ce qui a permis de fournir une évaluation dans un cadre homogène de toutes ces connaissances. Nous avons ainsi montré, contrairement aux travaux de l'état de l'art, l'impact positif et tranché des informations linguistiques en particulier morphologiques et sémantiques. Plutôt que de présupposer la pertinence de l'exploitation massive de toutes

ces informations, nous avons exposé un certain nombre de questions fondamentales sur leur complémentarité ou leur redondance dans leur quête de documents pertinents. Une analyse originale des corrélations existant entre elles nous a permis d'y répondre clairement. Nous avons en particulier mis au jour l'intérêt du couplage de connaissances morphologiques ou sémantiques mono-niveau, mais également de celui de connaissances bi-niveaux (morpho-sémantiques) qui, même si le nombre de cas sur lequel il est possible de l'exploiter est encore limité, est avéré.

Pour tirer parti de ce résultat fondamental, il convient de savoir comment combiner au mieux les connaissances linguistiques en RI et ce, de manière automatique. Nous avons pour cela proposé une méthode d'apprentissage supervisé basée sur un réseau de neurones qui fusionne les listes de résultats fournies par les différents index linguistiques et qui le fait en s'adaptant seule aux caractéristiques des requêtes traitées. Nous avons à travers nos expérimentations prouvé le gain effectif du couplage d'informations linguistiques en démontrant que notre système améliorait, en les rendant plus stables, les résultats obtenus par le meilleur des index pris individuellement.

Enfin, nous avons cherché également à améliorer la qualité des informations morphologiques qui se sont, tant dans nos travaux que dans ceux de l'état de l'art, dégagés comme étant actuellement les plus porteurs pour accroître les performances des SRI. Nous avons pour cela construit une nouvelle approche de détection de variantes morphologiques, simple et souple, et particulièrement bien adaptée aux contraintes de la RI. Elle ne nécessite ainsi aucune ressource externe, est entièrement automatique et peut être appliquée à différentes langues. L'emploi de cette technique en extension de requêtes a permis d'obtenir des résultats d'amélioration des capacités des SRI encore plus tranchés et beaucoup plus stables. Par le biais de cette méthode de détection de variantes morphologiques, nous démontrons ainsi qu'en construisant des outils plus souples, exploitant des informations linguistiques acquises en corpus directement à partir des collections, les techniques du TAL ont bien leur place en RI. Nos différentes contributions rapidement résumées ici montrent donc, qu'en exploitant autrement les informations linguistiques que ce soit en les combinant ou en les extrayant à partir de méthodes plus adaptées, on peut prétendre à des résultats plus catégoriques quant à leur intérêt en RI.

Points faibles et améliorations envisagées

Nos travaux ne sont toutefois pas exempts de certains points faibles. Parmi ceux-ci, nous pouvons évoquer tout d'abord l'aspect « boîte noire » du réseau de neurones retenu pour réaliser la tâche d'apprentissage sur laquelle s'appuie notre méthode de fusion de listes. Même si cela conduit à des résultats positifs, il est dommage que nous ne puissions savoir exactement parmi les différentes informations qui ont été combinées, celles qui ont contribué le plus fortement à l'amélioration des performances. L'utilisation d'une méthode d'apprentissage symbolique, capable d'obtenir les mêmes résultats mais de proposer également des éléments pour leur interprétation, constituerait une valeur ajoutée à notre approche. Sans changer de méthode d'apprentissage, nous pouvons également envisager d'évaluer notre méthode

de fusion à l'aide d'une technique particulière de validation du classifieur qui consiste à supprimer, pour chaque phase de test, une information linguistique particulière au sein des données d'entrée fournie au système d'apprentissage. En évaluant les performances du classifieur construit sans cette information mais avec toutes les autres et en procédant de la sorte pour chacune des connaissances, on pourrait déterminer, à travers les différences de résultats, les informations qui ont obtenu les meilleures performances et qui par conséquent ont le plus d'impact dans le couplage. Une autre faiblesse de notre système d'apprentissage est, comme nous l'avons souligné, qu'il ne fournit actuellement qu'une décision binaire sur la pertinence des documents. Cette contrainte ne nous permet pas, par conséquent, d'utiliser les mesures traditionnelles de RI, comme la MAP, et de comparer efficacement nos résultats à ceux obtenus par d'autres systèmes. Une amélioration du classifieur, qui consiste à donner un score aux documents en fonction de leur pertinence, est à l'étude et devrait par conséquent nous permettre de résoudre ce problème. De plus, à la suite des résultats obtenus lors de l'évaluation de l'efficacité de notre méthode de fusion, nous avons constaté que les gains observés correspondaient non pas à une hausse des taux de précision ou de rappel pris individuellement, mais à l'obtention d'un meilleur compromis entre ces deux taux. Ce constat nécessiterait d'être exploré plus en détail et pose également de nouvelles questions autour de l'évaluation des systèmes. Un autre bémol à nos travaux concerne le fait que nous n'avons actuellement réalisé nos expériences des chapitres 3 et 4 que sur une seule collection de documents. Afin de valider tous les résultats observés, il est indispensable de réitérer ces expérimentations sur d'autres types de collections et sur d'autres langues. Une autre limite de notre approche est liée à la qualité même des connaissances linguistiques utilisées dans notre travail. Tout en respectant la contrainte que nous nous étions fixée concernant la nécessité d'exploiter des informations linguistiques traditionnellement utilisées en RI, nous aurions pu chercher à améliorer leur qualité, par exemple en paramétrant de façon optimale le module de désambiguïsation des termes utilisé[6] ou en effectuant quelques traitements complémentaires, comme procéder à une racinisation (ou lemmatisation) des synonymes extraits... Le choix d'exploiter des informations sémantiques essentiellement obtenues à partir de méthodes basées sur des ressources (*e.g.* WORDNET) est aussi contestable, notamment de par leur couverture non exhaustive, et de par le problème de l'adéquation des informations contenues avec le domaine de la collection. Comme nous l'avions fait pour les connaissances syntaxiques, le couplage de ces connaissances sémantiques avec d'autres — acquises à l'aide de méthodes différentes — (comme le recours à des informations de cooccurrences ou de similarité de cooccurrents) aurait été intéressant à évaluer.

[6] Comme nous l'avions souligné, nous avions gardé les paramètres par défaut afin de ne pas augmenter les temps de traitement.

Perspectives

Les différentes pistes explorées dans le cadre de cette thèse nous ont amenée à envisager des perspectives intéressantes. Nous présentons ici celles qui nous paraissent les plus prometteuses.

Lorsque nous avons étudié l'intérêt de coupler des informations linguistiques multi-niveaux en RI, nous avons observé des différences importantes entre les connaissances quant à leur efficacité respective à améliorer les performances des systèmes. Nous avons plus particulièrement mis l'accent sur le faible apport des informations de nature syntaxique et de certaines connaissances d'ordre sémantique. Une perspective à court terme est de proposer de nouvelles expérimentations en exploitant des connaissances plus pertinentes d'un point de vue linguistique, mêmes si celles-ci sont plus complexes à acquérir. En effet, ces expériences nous permettrait d'évaluer précisément si les résultats obtenus sont liés à la qualité des représentations proposées ou à la façon de les utiliser en RI. Parmi les informations susceptibles de mieux représenter le contenu textuel des documents et requêtes, nous pensons plus particulièrement à des informations sémantiques acquises en corpus telles que le recours aux liens nom-verbe qui se sont révélés pertinents en RI pour l'expansion de requêtes (Claveau et Sébillot, 2004b; Grefenstette, 1997) et qui pourraient également être exploités lors de l'indexation. À un niveau syntaxique, des informations d'entités nommées de qualité ou des connaissances de positionnement de mots (ordre, distance) pourraient aussi certainement améliorer les résultats obtenus.

À partir de nos travaux sur la détection automatique de variantes morphologiques qui ont permis de faire ressortir l'intérêt en RI de techniques souples et auto-adaptées, nous pourrions envisager de transformer notre méthode pour acquérir d'autres types de variantes linguistiques. Les travaux de Claveau et L'Homme (2005a) ont déjà montré l'efficacité de cette méthode pour détecter des mots en relations morpho-sémantiques. Sur le plan syntaxique, nous pourrions envisager de l'appliquer pour le repérage de structures syntaxiques équivalentes (telles que celles étudiées par Jones et Tait (1984)).

Enfin, dans l'ensemble de nos travaux, nous nous sommes appuyée sur l'hypothèse que si la convergence entre le TAL et la RI ne donnait pas des résultats plus tranchés, cela était dû principalement au fait que les traitements linguistiques n'étaient pas adaptés au domaine de la RI. Une autre façon d'envisager le problème serait de considérer que ce sont les mécanismes de représentation des contenus textuels et de mise en correspondance tels qu'ils sont proposés par la RI qui ne pas assez souples pour exploiter pleinement la richesse des informations linguistiques. En effet, comme nous l'avons remarqué dans le premier chapitre, les modèles de RI sont, pour beaucoup, limités dans leur façon de représenter le contenu des documents et requêtes. Ces représentations (en « sac de mots » pour modèle vectoriel) ne prennent en compte ni l'ordre des mots ni les dépendances entre les termes et sont par conséquent peu adaptés à accueillir des informations plus riches que de simples mots. Une perspective particulièrement intéressante à notre travail serait de ré-étudier l'apport du couplage d'informations linguistiques sur d'autres modèles de RI plus aptes à prendre en compte nos représentations enrichies. Nous pensons plus particulièrement aux modèles de langue qui proposent des solutions prometteuses pour l'intégration de connaissances linguistiques.

Annexe A

Caractéristiques de la collection TIPSTER

Les documents, requêtes et fichiers de pertinence sur lesquels nous nous appuyons pour la plupart de nos expérimentations (sauf pour quelques-unes du chapitre 5) proviennent de la collection TIPSTER (présentée en section 3.4.1) utilisée lors des campagnes d'évaluation TREC. Nous exploitons plus précisément un sous-ensemble de cette collection qui regroupe des articles de journaux issus du *Wall Street Journal* des années 1986 à 1992.

Nous présentons tout d'abord quelques statistiques (issues des données officielles de TREC) sur le contenu de cette sous-collection. À titre illustratif, nous donnons ensuite un exemple de requête et document la composant.

Statistiques sur la collection *Wall Street Journal*

Le tableau A.1 présente quelques indications des caractéristiques de la collection (répartie en 2 CD-ROM).

Caractéristiques	Répartition	Collection *Wall Street Journal*
Taille de la collection (*megabytes*)	CD 1	295
	CD 2	255
	Total	550
Nombre de documents	CD1	98736
	CD2	74520
	Total	173256
Nombre moyen de termes par document	CD1	329
	CD2	377
Nombre médian de termes par document	CD1	182
	CD2	218
Nombre total de termes différents	CD1	156298
	CD2	153725
Nombre total de termes différents et apparaissant une seule fois	CD1	64656
	CD2	64844
Nombre total de termes différents et apparaissant plus d'une fois	CD1	91642
	CD2	88881
Nombre moyen d'occurrences supérieures à 1	CD1	199
	CD2	178

TAB. A.1 – Quelques données sur la collection du *Wall Street Journal*

Exemples de documents et requêtes

Les documents et requêtes de la collection sont au format SGML. Un traitement de suppression des balises est au préalable effectué. Les documents de cette collection (articles de journaux) sont décomposés en un certain nombre de champs (identifiant, date, titre...). Dans nos expérimentations, seul le contenu informationnel du document nous intéresse. Nous retenons donc uniquement l'identifiant du document (champ <DOCNO>) et le corps même du texte (champ <TEXT>).

Exemple de document

L'exemple suivant présente un article issu de la collection du *Wall Street Journal* :

```
<DOC>
<DOCNO> WSJ870320-0197 </DOCNO>
<HL> Rothmans of Canada
Considers Making
A Large Acquisition</HL>
<DD> 03/20/87</DD>
<SO> WALL STREET JOURNAL (J)</SO>
<IN> T.ROC
```

```
TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)
FOOD &amp;amp ; BEVERAGE, HOUSEHOLD GOODS, SUPERMARKETS, TOBACCO (FAB) </IN>
<DATELINE> TORONTO </DATELINE>
<TEXT>
The Canadian unit of Rothmans International PLC said it would consider
spending as much as $500 million (Canadian) on an acquisition. This would
be "about double" the $250 million in net proceeds it expects to receive
from two recently announced transactions.
They include the December merger of Rothmans's tobacco business with
a unit of Philip Morris Cos., New York, and the proposed sale of
Rothmans's 50.1% stake in Canada's third largest brewery, Carling
O'Keefe Ltd.
In a presentation to analysts, Rothmans Inc.'s president and chief
executive officer, Patrick Fennell, said the bid for Rothmans's Carling
shares, made last month by a unit of Australia-based Elder's IXL Ltd.,
would net Rothmans about $180 million, if approved, as expected, by
governmental and regulatory authorities. The Elders offer is scheduled
to expire next Thursday.
Mr. Fennell said Rothmans has "no targets" currently, and it hasn't
retained a financial adviser to seek acquisitions. But he said the
tobacco company is interested in "opportunities in consumer packaged
goods," which he described as "largely recession-proof" and "an offset
to the business we're clearly dependent on here in Canada."
Mr. Fennell said the merger of Rothmans's tobacco business with that of
Benson &amp;amp ; Hedges (Canada) Inc. to form Rothmans, Benson &amp;amp ; Hedges,
gives the new company a 32% share of the Canadian cigarette market,
second to the 51% share held by market leader Imperial Tobacco Ltd., a
unit of Montreal's Imasco Ltd.
Rothmans Inc. is 71.2%-owned by London-based Rothmans International.
</TEXT>
</DOC>
```

Exemple de requête

Les requêtes sont plus précisément des *topics* (*cf.* présentation section 1.4.1). Un *topic*, dont un exemple est présenté ci-dessous, est considéré comme une description enrichie du besoin d'information de l'utilisateur. Il se compose de différents champs. Le champ <top> contient l'ensemble de la question dont l'identifiant est donné par le champ <num>. Le champ <dom> indique le domaine général de la question ; le champ <title> représente le titre de la question ; le champ <desc> est une description de la question (2-3 phrases). Le champ <narr> précise les documents qui sont attendus en réponse ; le champ <con> propose une liste de concepts proches de la question. Le champ <fac> permet de spécifier des éléments du texte susceptibles d'être importants

(comme la nationalité (champ <nat>) dans l'exemple présenté) et champ <def> offre la possibilité d'associer des définitions aux concepts.

Pour nos expériences, nous exploitons seulement les champs <num>, <title> et <desc>.

```
<top>
<head> Tipster Topic Description</head>
<num> Number : 066</num>
<dom> Domain : Science and Technology</dom>
<title> Topic : Natural Language Processing</title>
<desc> Description : Document will identify a type of natural language
processing technology which is being developed or marketed in the
U.S.</desc>
<narr> Narrative : A relevant document will identify a company or
institution developing or marketing a natural language processing
technology, identify the technology, and identify one or more features
of the company's product.</narr>
<con> Concept(s) :
1. natural language processing
2. translation, language, dictionary, font
3. software applications
</con>
<fac> Factor(s) :
<nat> Nationality : U.S.</nat>
</fac>
<def> Definition(s) :</def>
</top>
```

Annexe B

Analyse linguistique des documents et requêtes

Nous nous intéressons ici aux informations linguistiques extraites automatiquement à partir des textes par le biais d'outils et de techniques du TAL[1]. Nous donnons l'exemple d'un document (issu de notre collection et correspondant plus précisément ici à un paragraphe) sur lequel ont été appliqués 11 traitements linguistiques différents. Nous présentons le document tel qu'il a été obtenu après avoir été traité par chacun des 11 outils utilisés[2].

Document original

À titre de comparaison, nous présentons le texte original tel qu'il est présent dans la collection (sans analyse linguistique) :

American Express Co.'s board will be asked Monday to approve a public offering of part of the company's Shearson Lehman Brothers Inc. securities unit, company sources said. Meanwhile, American Express announced that it had reached an "understanding" to sell a 13% stake in Shearson to Tokyo-based Nippon Life Insurance Co. for about $530 million. American Express sources said that if the board approves the public offering, an announcement would be made late Monday, followed by a news conference Tuesday. The sources declined to say how much of Shearson, one of the nation's largest and most profitable securities firms, might be offered to the public. Board approval is also necessary before a definitive agreement can be reached withlife insurance by far Japan's largest life insurance concern. The Nippon Life accord also hinges on the approval of Japan's Ministry of Finance.

[1] Pour une description de ces outils, se référer à la section 3.3.1

[2] Nous présentons ici les données brutes, c'est-à-dire correspondant directement au résultat fourni en sortie d'un outil (*i.e.* sans avoir modifié la casse, la ponctuation...).

Lemmes

Après un traitement de lemmatisation, c'est-à-dire conversion de toutes les formes des mots en leur lemme, le document obtenu est le suivant :

American Express Co. 's board will be ask Monday to approve a public offering of part of the company 's Shearson Lehman Brothers Inc. security unit , company source say . meanwhile , American Express announce that it have reach an "understanding" to sell a 13 % stake in Shearson to Tokyo-based Nippon Life Insurance Co. for about $ 530 million . American Express source say that if the board approve the public offering , an announcement would be make late Monday , follow by a news conference Tuesday . the source decline to say how much of Shearson , one of the nation 's large and most profitable security firm , might be offer to the public . board approval be also necessary before a definitive agreement can be reach with Nippon Life , by far Japan 's large life insurance concern . the Nippon Life accord also hinge on the approval of Japan 's Ministry of Finance .

Racines

Après racinisation, le document correspond à un ensemble de racines :

american express co.'s board will be ask mondai to approv a public offer of part of the compani's shearson lehman brother inc. secur unit, compani sourc said. meanwhil, american express announc that it had reach an "understand" to sell a 13% stake in shearson to tokyo-base nippon life insur co. for about $530 million. american express sourc said that if the board approv the public offer, an announc would be made late mondai, follow by a new confer tuesdai. the sourc declin to sai how much of shearson, on of the nation's largest and most profit secur firm, might be offer to the public. board approv is also necessari befor a definit agreement can be reach with nippon life, by far japan's largest life insur concern. the nippon life accord also hing on the approv of japan's ministri of financ.

Termes simples + étiquettes grammaticales

On procède à une analyse morpho-syntaxique des documents ; une étiquette grammaticale est donc associée aux termes :

American_NP Express_NP Co_NP s_POS board_NN will_MD be_VB asked_VVN Monday_NP to_TO approve_VV a_DT public_JJ offering_NN of_IN part_NN of_IN the_DT company_NN s_POS Shearson_NP Lehman_NP Brothers_NPS Inc_NP securities_NNS unit_NN company_NN sources_NNS said_VVD Meanwhile_RB American_NP Express_NP announced_VVD that_IN it_PP had_VHD reached_VVN an_DT "_ understanding_NN "_ to_TO sell_VV a_DT 13_CD

%_NN stake_NN in_IN Shearson_NP to_TO Tokyo-based_JJ Nippon_NP Life_NP Insurance_NP Co_NP for_IN about_RB $_ 530_CD million_CD American_NP Express_NP sources_NNS said_VVD that_IN if_IN the_DT board_NN approves_VVZ the_DT public_JJ offering_NN an_DT announcement_NN would_MD be_VB made_VVN late_JJ Monday_NP followed_VVN by_IN a_DT news_NN conference_NN Tuesday_NP The_DT sources_NNS declined_VVD to_TO say_VV how_WRB much_JJ of_IN Shearson_NP one_CD of_IN the_DT nation_NN s_POS largest_JJS and_CC most_RBS profitable_JJ securities_NNS firms_NNS might_MD be_VB offered_VVN to_TO the_DT public_NN Board_NN approval_NN is_VBZ also_RB necessary_JJ before_IN a_DT definitive_JJ agreement_NN can_MD be_VB reached_VVN with_IN Nippon_NP Life_NP by_IN far_RB Japan_NP s_POS largest_JJS life_NN insurance_NN concern_NN The_DT Nippon_NP Life_NP accord_NN also_RB hinges_VVZ on_IN the_DT approval_NN of_IN Japan_NP s_POS Ministry_NP of_IN Finance_NP

Termes complexes

Nous représentons ici le texte au travers des termes complexes qui le composent (les constituants sont groupés par un caractère _) :

%_stake Board_approval Company_source Definitive_agreement Express_source Inc_security Insurance_company Japanese_company Japanese_individual Life_accord Life_insurance Life-Shearson_partnership Major_security Ministry_finance
News_conference Nippon_life Public_offering

Bigrammes

La description du texte par le biais d'informations de bigrammes[3] est :

Nippon_Life American_Express security_firm life_insurance public_offering Express_source investing_Nippon Life_Shearson large_life Tokyo_based board_approval major_security Insurance_Co definitive_agreement insurance_company news_conference Inc_security Shearson_Lehman source_decline Lehman_Brothers conference_Tuesday about_million Life_Insurance ask_Monday Brothers_Inc

Trigrammes

[3]L'ordre des mots est modifié puisque l'outil utilisé trie les bigrammes dans l'ordre décroissant de leur fréquence dans la collection.

La description d'un document à l'aide d'informations de trigrammes[4] donne :

American_Express_source large_security_firm Life_American_Express
Tokyo_based_Nippon large_life_insurance Inc_security_unit security_unit_company meanwhile_American_Express American_Express_announce Nippon_Life_Shearson public_board_approval Nissei_Life_America million_American_Express
American_Express_Co Nippon_Life_association Life_Insurance_Co unit_company_source American_Express_decline Brothers_Inc_security life_insurance_concern
news_conference_Tuesday

Groupes nominaux

Le document vu comme un ensemble de syntagmes nominaux correspond à :

American_Express_Co. public_offering Shearson_Lehman_Brothers_Inc._security_unit company_source American_Express Tokyo-based_Nippon_Life_Insurance_Co.
American_Express_source public_offering late_Monday news_conference source_decline
large_and__profitable_security_firm board_approval definitive_agreement Nippon_Life large_life_insurance_concern Nippon_Life_accord American_Express_decline

Noms propres

Le document représente les termes qui ont été identifiés comme noms propres :

American Express Co Monday Inc American Express Shearson Nippon Life Insurance Co American Express Monday Tuesday Shearson Nippon Life Japan Nippon Life accord Japan Ministry Finance

Termes simples + étiquettes sémantiques

Après désambiguïsation, une étiquette sémantique issue de WORDNET est associée à chaque terme du document. La lettre après le terme correspond à sa catégorie grammaticale (un nom n, un verbe v...). Le chiffre représente le numéro de sens du terme dans WORDNET.

American#n#3 Express#v#2 Co#n#4 board#n#2 will#n#3 be#v#1 ask#v#7 Monday#n#1 approve#v#1 public#a#1 offering#n#1 part#n#4 company#n#1 Brother#n#2 Inc#n#1 security#n#2 unit#n#3 company#n#1 source#n#1

[4] Même remarque que pour les bigrammes pour le changement dans l'ordre des mots.

say#v#8 Meanwhile#r#2 American#n#1 Express#v#2 announce#v#1 have#v#15 reach#v#1 understanding#n#2 sell#v#1 stake#n#1 Nippon#n#1 Life#n#5 Insurance#n#1 Co#n#4 million#a#1 American#n#1 Express#v#6 source#n#1 say#v#1 board#n#2 approve#v#1 public#a#1 offering#n#1 announcement#n#2 be#v#1 make#v#33 late#a#5 Monday#n#1 follow#v#12 news#n#4 conference#n#1 Tuesday#n#1 source#n#1 decline#v#3 say#v#8 how#r#2 much#a#1 one#n#1 nation#n#1 large#a#1 profitable#a#1 security#n#2 firm#n#1 might#n#1 be#v#1 offer#v#2 public#a#1 Board#n#2 approval#n#4 be#v#1 also#r#1 necessary#a#1 definitive#a#3 agreement#n#1 can#n#5 be#v#1 reach#v#6 Nippon#n#1 Life#n#5 far#a#1 Japan#n#2 large#a#1 life#n#5 insurance#n#1 concern#n#3 Nippon#n#1 Life#n#5 accord#n#1 also#r#1 hinge#n#1 approval#n#4 Japan#n#2 Ministry#n#3 Finance#n#1

Termes simples + synonymes

Un ensemble de synonymes, indiqués entre parenthèses, est associé à chaque terme du document :

American (American) Express (express, verbalize, verbalise, utter, give_tongue_to) Co (Colorado, Centennial_State, CO) board (board) will (will, testament) be (be) ask (necessitate, ask, postulate, need, require, take, involve, call_for, demand) Monday (Monday, Mon) approve (approve, O.K., okay, sanction) public (public) offering (offer, offering) part (part, portion) company (company) Brother (brother)
Inc (Iraqi_National_Congress, INC) security (security, certificate) unit (unit, social_unit) company (company) source (beginning, origin, root, rootage, source) say (pronounce, articulate, enounce, sound_out, enunciate say)
Meanwhile (meanwhile, meantime, in_the_meantime) American (American) Express (express, verbalize, verbalise, utter, give_tongue_to) announce (announcen denote) have (have, have_got, hold) reach (reach, make, attain, hit, arrive_at, gain) understanding (agreement, understanding) sell (sell) stake (interest, stake) Nippon (Japan, Nippon, Nihon) Life (life, lifetime, lifespan) Insurance (insurance) Co (Colorado, Centennial_State, CO) million (million, a_million) American (American) Express (press_out, express, extract) source (beginning, origin, root, rootage, source) say (state, say, tell) board (board) approve (approve, O.K., okay, sanction) public (public) offering (offer, offering) announcement (announcement, promulgation) be (be) make (make, create) late (late) Monday (Monday, Mon) follow (follow) news (news) conference (conference) Tuesday (Tuesday, Tues) source (beginning, origin, root, rootage, source) decline (refuse, decline) say (pronounce, articulate, enounce, sound_out, enunciate say) how (how, however) much (much) one (one, 1, I, ace, single, unity) nation (state, nation, country, land, commonwealth, res_publica, body_politic) large (large ,big) profitable (profitable) security (security, certificate) firm (firm, house, business_firm) might (might, mightiness, power) be (be) offer (offer, proffer) public (public) Board (board) approval (approval, commendation) be (be) also (besides, too, also, likewise, as_well) necessary (necessary) definitive (definitive, determinate) agreement (agreement, understanding) can (toilet,

can, commode, crapper, pot, potty, stool, throne) be (be) reach (reach, extend_to touch) Nippon(Japan, Nippon Nihon) Life (life, lifetime, lifespan) far (far) Japan (Japan, Nippon, Nihon) large (large, big) life (life, lifetime, lifespan) insurance (insurance) concern (business, concern, business_concern, business_organization, business_organisation) Nippon (Japan, Nippon, Nihon) Life (life, lifetime, lifespan) accord (agreement, accord) also (besides, too, also, likewise, as_well) hinge (hinge, flexible_joint) approval (approval, commendation) Japan (Japan, Nippon, Nihon) Ministry (ministry) Finance (finance)

Mots reliés

Dans ce document, certains termes reçoivent un ensemble de mots reliés morphosémantiquement :

American Express (expression) board will (bequeath will leave) ask Monday approve (approbation, approver, blessing, approval, approving) offering (offer) company Brother Inc security unit (unitize, unitise) source say Meanwhile American (Americanize, Americanise) announce (announcer, announcer) reach understanding sell (sell, seller, marketer, vender, vendor, trafficker, selling, merchandising, marketing) Nippon Life Insurance (cover, insure, underwrite) million American (Americanize, Americanise) source say (say) approve (approbation, approver, blessing, approval, approving) offering (offer) be make (devising, fashioning, making) Monday follow (follower, pursuit, chase, following) conference (confer, confabulate, confab, consult) source decline say how much one nation large profitable security firm might be offer (offer, offering, offerer, offeror) Board approval (approbate) also necessary definitive agreement (agree) be reach (range, reach) Life far Japan large life insurance (cover, insure, underwrite) Nippon Life accord also hinge (hinge) Japan Ministry Finance (finance, finance)

Bibliographie

Alvarez, C., Langlais, P. et Nie, J.-Y. (2003). Word Pairs in Language Modeling for Information Retrieval. Rapport interne, Laboratoire de recherche appliquée en linguistique informatique (RALI), Montréal, Canada.

Arampatzis, A., Tsoris, T. et Koster, C. H. (1996). irena : Information Retrieval Engine Based on Natural Language Analysis. Rapport technique, Computing Science Institute, Nijmegen, Pays-Bas.

Arampatzis, A., Weide, T. v., Koster, C. H. et Bommel, P. v. (2000). Linguistically Motivated Information Retrieval. *In* Kent, A., éditeur : *Encyclopedia of Library and Information Science*, pages 201–222. M. Dekker, New York, NY Basel.

Aslam, J. et Montague, M. (2001). Models for Metasearch. *In Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, New-Orléans, États-Unis.

Audibert, L. (2003). *Outils d'exploration de corpus et désambiguïsation lexicale automatique*. Thèse de doctorat, Université d'Aix-Marseille I - Université de Provence, Aix-en-Provence, France.

Baeza-Yates, R. A. et Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press - Addison-Wesley.

Banerjee, S. et Pedersen, T. (2003). Extended Gloss Overlap as a Measure of Semantic Relatedness. *In Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexique.

Bar-Ilan, J., Levene, M. et Mat-Hassan, M. (2004). Dynamics of Search Engine Rankings - A Case Study. *In Proceedings of the 3rd International Workshop on Web Dynamics*, New-York,États-Unis.

Bartell, B. T., Cottrell, G. W. et Belew, R. K. (1994). Automatic Combination of Multiple Ranked Retrieval Systems. *In Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Dublin, Irlande.

Baziz, M. (2005). *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. A., Frieder, O. et Goharian, N. (2004). Fusion of Effective Retrieval Strategies in the same Information Retrieval System. *Journal of the American Society for Information Science and Technology*, 55(10):859–868.

Belkin, N. J., Kantor, P., Fox, E. A. et Shaw, J. A. (1995). Combining Evidence of Multiple Query Representations for Information Retrieval. *Information Processing and Management*, 31(3):414–448.

Berrut, C. (1990). Indexing Medical Reports - The rime Approach. *Information Processing and Management*, 26(1):93–109.

Besançon, R. (2002). *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de textes : application au calcul de similarités sémantiques dans le cadre du modèle DSIR*. Thèse de doctorat, École polytechnique fédérale de Lausanne, Lausanne, Suisse.

Bilotti, M. W., Katz, B. et Lin, J. (2004). What Works Better for Question Answering : Stemming or Morphological Query Expansion ? *In Proceedings of Question Answering Workshop : ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Sheffield, Angleterre.

Bommier-Pincemin, B. (1999). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. Thèse de doctorat, Université de la Sorbonne, Paris, France.

Bookstein, A., Klein, A. T. et Raita, T. (1995). Content-Bearing Words by Serial Clustering. *In Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Seattle, États-Unis.

Borda, C. d. (1781). Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences.

Boughanem, M. (1992). *Système de recherche d'informations : d'un modèle classique à un modèle connexionniste*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Boughanem, M. (2003). Outils de validation en recherche d'information. La campagne d'évaluation TREC. http ://inforsid2003.loria.fr/resumeConfRI.pdf.

Boughanem, M., Dkaki, T., Mothe, J. et Soulé-Dupuy, C. (1998). mercure at TREC-7. *In Proceedings of the 7th International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

Brill, É. (1992). A Simple Rule-Based Part-of-Speech Tagger. *In Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP)*, Trento, Italie.

Buckley, C., Salton, G. et Allan, J. (1992). Automatic Retrieval with Locality Information using smart. *In Proceedings of the 1st International Conference on Text Retrieval (TREC)*, Maryland, États-Unis.

Buckley, C. et Voorhees, E. M. (2000). Evaluating Evaluation Measure Stability. *In Proceedings of the 23th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Athènes, Grèce.

Callan, J. P., Croft, W. B. et Harding, S. M. (1992). The inquery Retrieval System. *In Proceedings of the International Conference on Database and Expert Systems Applications*, Valence, Espagne.

Carlberger, J., Dalianis, H., Hassel, M. et Knutsson, O. (2001). Improving Precision in Information Retrieval for Swedish using Stemming. *In Proceedings of the 13th Nordic Conference on Computational Linguistics (NODALIDA)*, Uppsala, Suède.

Chevallet, J.-P. (1992). *Un modèle logique de recherche d'information appliqué au formalisme des graphes conceptuels. Le prototype* elen *et son expérimentation sur un corpus de composants logiciels.* Thèse de doctorat, Université Joseph Fourier, Grenoble, France.

Chevallet, J.-P. (2004). Modélisation logique pour la recherche d'information. *In* Lavoisier, éditeur : *Les systèmes de recherche d'information*, pages 105–138. Hermes.

Claveau, V. (2003). *Acquisition automatique de lexiques sémantiques pour la recherche d'information.* Thèse de doctorat, Université de Rennes 1, Rennes, France.

Claveau, V. et L'Homme, M.-C. (2005a). Apprentissage par analogie de liens sémantiques entre dérivés morphologiques. *In Proceedings of conférence de terminologie et intelligence artificielle (TIA)*, Rouen, France.

Claveau, V. et L'Homme, M.-C. (2005b). Structuring Terminology by Analogy Machine Learning. *In Proceedings of the International Conference on Terminology and Knowledge Engineering (TKE)*, Copenhague, Danemark.

Claveau, V. et Sébillot, P. (2004a). Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe. *Traitement automatique des langues*, 45(1):153–182.

Claveau, V. et Sébillot, P. (2004b). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. *In Proceedings of 11ème conférence annuelle sur le traitement automatique des langues naturelles (TALN)*, Fès, Maroc.

Cleverdon, C. W. (1967). The Cranfield Tests on English Language Devices. *Aslib Proceedings*, 19(6):173–184.

Condorcet, M. d. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.

Cornuéjols, A. et Miclet, L. (2002). *Apprentissage artificiel. Concepts et algorithmes*. Eyrolles.

Croft, W. B. (1997). Combining Approaches to Information Retrieval. *In* Croft, W. B., éditeur : *Advances in Information Retrieval : Recent Research from the Center for Intelligent Information Retrieval*, pages 1–36. Kluwer Academic Publishers.

Croft, W. B. et Harper, D. J. (1979). Using Probabilistic Models of Document Retrieval without Relevance Information. *Journal of Documentation*, 35(4):285–295.

Cronen-Townsend, S., Zhou, Y. et Croft, W. B. (2002). Predicting Query Performance. *In Proceedings of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Tampere, Finlande.

Cunningham, H. (2002). gate, a General Architecture for Text Engineering. *Computers and the Humanities*, 36(1):223–254.

Daille, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In* Resnik, P. et Klavans, J., éditeurs : *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press.

Daille, B. (2002). *Découvertes linguistiques en corpus*. Mémoire d'habilitation à diriger des recherches, Université de Nantes, Nantes, France.

Daille, B., Fabre, C. et Sébillot, P. (2002). Applications of Computational Morphology. *In* Boucher, P., éditeur : *Many Morphologies*, pages 210–234. Cascadilla Press, Somerville.

Deerwester, S., Dumais, S. T., Furnas, T. K. et Harman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Delafosse, L. (1999). Présentation du TAL. http ://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/cours/parcours/introtal.htm.

Denis, F. et Gilleron, R. (2000). Apprentissage à partir d'exemples. Notes de cours. http ://www.grappa.univ-lille3.fr/polys/apprentissage/.

Dillon, G. M. et Gray, A. S. (1983). fasit : A Fully Automatic Syntactically Based Indexing System. *Journal of the American Society for Information Science*, 34(2):99–108.

Dubois, D., Prade, H. et Yager, R. R. (1997). *Fuzzy Information Engineering - A Guided Tour of Applications*. Wiley Computer Publishing.

Dumais, S. T. (1991). Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

FAGAN, J. (1987). *Experiments in Automatic Phrase Indexing for Document Retrieval : A Comparison of Syntactic and Non-Syntactic Methods.* Thèse de doctorat, Université de Cornell, New-York, États-Unis.

FAGIN, R., KUMAR, R. et SIVAKUMAR, D. (2003). Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160.

FAJAH, N., GODIN, R., MISSAOUI, R., DAVID, S. et PLANTE, P. (1996). Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte. *Canadian Journal of Information and Library Science*, 21(1):1–21.

FELLBAUM, C. (1998). WORDNET *: An Electronic Lexical Database.* C. Fellbaum, The MIT Press, Cambridge, Massachussets, États-Unis.

FIESLER, E. et BEALE, R. (1996). *Handbook of Neural Computation.* Institute of Physics and Oxford University Press.

FOX, E. A. (1983). *Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types.* Thèse de doctorat, Université de Cornell, New-York, États-Unis.

FOX, E. A. et SHAW, J. A. (1994). Combination of Multiple Searches. *In Proceedings of the 2nd International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

FRAKES, W. B. (1992). Stemming Algorithms. *In* FRAKES, W. B. et BAEZA-YATES, R., éditeurs : *Information Retrieval : Data Structures and Algorithms*, pages 131–160. Prentice Hall.

FREUND, G. E. et WILLETT, P. (1982). Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure. *Information Technology : Research and Development*, 1:177–187.

FRIBURGER, N. et MAUREL, D. (2002). Textual Similarity Based on Proper Names. *In Proceedings of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Tampere, Finlande.

FUHR, N. (2005). Information Retrieval - From Information Access to Contextual Retrieval. *In* EIBL, M., WOLFF, C. et WOMSER-HACKER, C., éditeurs : *Designing Information Systems. Festschrift für Jürgen Krause*, pages 47–57. UVK Verlagsgesellschaft.

FULLER, M. et ZOBEL, J. (1998). Conflation-Based Comparison of Stemming Algorithms. *In Proceedings of the 3th Australian Document Computing Symposium*, Sydney, Australie.

GAUCH, S., WANG, J. et RACHAKONDA, S. M. (1999). A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases. *ACM Transactions on Information Systems*, 17(3):250–269.

GAUSSIER, ., GREFENSTETTE, G., HULL, D. et ROUX, R. (2000). Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues*, 41(2):473–493.

GAUSSIER, ., GREFENSTETTE, G. et SCHULZE, M. (1997). Traitement du langage naturel et recherche d'informations : quelques expériences sur le français. *In Proceedings of 1ères journées scientifiques et techniques du réseau francophone de l'ingénierie de la langue de l'AUPELF-UREF*, Avignon, France.

GAUSSIER, E. (1999). Unsupervised Learning of Derivational Morphology from Inflectional Corpora. *In Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, Maryland, États-Unis.

GOLDSMITH, J. A. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

GOLDSMITH, J. A., HIGGINS, D. et SOGLASNOVA, S. (2001). Automatic Language-Specific Stemming in Information Retrieval. *In Proceedings of Workshop of Cross-Language Evaluation Forum (CLEF)*, Lisbonne, Portugal.

GONZALO, J., VERDEJO, F., CHUGUR, I. et CIGARRAN, I. J. (1998). Indexing with WORDNET Synsets can Improve Text Retrieval. *In Proceedings of the Workshop on Usage of* WORDNET *for NLP, 17th International Conference on Computational Linguistics (COLING)*, Montréal, Canada.

GREFENSTETTE, G. (1994). Corpus-Derived First, Second and Third-Order Word Affinities. *In Proceedings of EURALEX International Congress*, Amsterdam, Pays-Bas.

GREFENSTETTE, G. (1997). SQLET : Short Query Linguistic Expansion Techniques : Palliating One-Word Queries by Providing Intermediate Structure to Text. *In Proceedings of 5ème conférence internationale sur la recherche d'informations assistée par ordinateur (RIAO)*, Montréal, Canada.

GRIVOLLA, J. (2001). Évaluation et prédiction des difficultés des requêtes dans la recherche documentaire pour l'optimisation des systèmes interactifs. Mémoire de DEA, Laboratoire informatique d'Avignon, Avignon, France.

GUARINO, N., MASSOLO, C. et VETERE, G. (1999). ONTOSEEK : Content-Based Access to the Web. *IEEE Intelligent Systems*, 14(3):70–80.

HADDAD, H. (2002). *Extraction et impact des connaissances sur les performances des systèmes de recherche d'information.* Thèse de doctorat, Université Joseph Fourier, Grenoble, France.

HADDAD, H. (2003). Utilisation des syntagmes nominaux dans un système de recherche d'information. *In Proceedings of 19èmes journées de bases de données avancées (BDA)*, Lyon, France.

HARMAN, D. (1991). How Effective is Suffixing ? *Journal of the American Society for Information Science*, 42(1):7–15.

HARMAN, D. (1992). Relevance Feedback Revisited. *In Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Copenhague, Danemark.

HATHOUT, N. (2001). Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. *In Proceedings of 8ème conférence annuelle sur le traitement automatique des langues naturelles (TALN)*, Tours, France.

HEARST, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *In Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Nantes, France.

HERTZ, J., KROGH, A. et PALMER, G., éditeurs (1991). *Introduction to the Theory of Neural Computation*. Addison Wesley.

HÉRAULT, J. et JUTTEN, C., éditeurs (1994). *Réseaux neuronaux et traitement du signal*. Hermès.

HULL, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. *In Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Pittsburgh, États-Unis.

HULL, D. (1996). Stemming Algorithms - A Case Study for Detailed Evaluation. *Journal of the American Society of Information Science*, 47(1):70–84.

HULL, D. et GREFENSTETTE, G. (1996). A Detailed Analysis of English Stemming Algorithms. Rapport technique, Xerox Research Centre Europe, Meylan, France.

HULL, D., GREFENSTETTE, G., SCHULZE, B. M., SCHÜTZE, H. et PEDERSEN, J. O. (1997). Xerox TREC-5 Site Report : Routing, Filtering, NLP and Spanish Tracks. *In Proceedings of the 5th Internation Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

JACQUEMIN, C. (1994). FASTR : A Unification-Based Front-End to Automatic Indexing. *In Proceedings of 4ème conférence internationale sur la recherche d'informations assistée par ordinateur (RIAO)*, New York, États-Unis.

JACQUEMIN, C., KLAVANS, J. L. et TZOUKERMANN, E. (1997). Expansion of Multi-Word Terms for Indexing and Retrieval using Morphology and Syntax. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, Madrid, Espagne.

JACQUEMIN, C. et ROYAUTÉ, J. (1994). Retrieving Terms and their Variants in a Lexicalised Unification-Based Framework. *In Proceedings of the 17th ACM International*

Conference on Research and Development in Information Retrieval (SIGIR), Dublin, Ireland.

Jing, Y. et Croft, W. B. (1994). An Association Thesaurus for Information Retrieval. *In Proceedings of 4ème conférence internationale sur la recherche d'informations assistée par ordinateur (RIAO)*, New York, États-Unis.

Jones, K. S. et Tait, J. I. (1984). Automatic Search of Term Variant Generation. *Journal of Documentation*, 1(1):50–66.

Jouis, C. (1995). seek, un logiciel d'acquisition de connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. *In Proceedings of 6èmes journées sur l'acquisition et la validation (JAVA)*, Grenoble, France.

Jourlin, P., Johnson, S. E., Spärck Jones, K. et Woodland, P. C. (2000). Spoken Document Representations for Probabilistic Retrieval. *Spoken Document Representations for Probabilistic Retrieval Speech Communication*, 32(1-2):21–36.

Khan, L. R. (2000). *Ontology-Based Information Selection.* Thèse de doctorat, Université de Californie du Sud, Los Angeles, États-Unis.

Khoo, C. S. G. (1995). *Automatic Identification of Causal Relations in Text and their Use for Improving Precision in Information Retrieval.* Thèse de doctorat, Université de Syracuse, New-York, États-Unis.

Kilgarriff, A. et Palmer, M. (2000). Special Issue on Senseval. *Computers and the Humanities*, 34(1/2).

Korhage, R. (1997). *Information Storage and Retrieval.* John Wiley and Sons.

Kraaij, W. et Pohlmann, R. (1996). Viewing Stemming as Recall Enhancement. *In Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Zürich, Suisse.

Kraft, D. H. et Buell, D. A. (1983). Fuzzy Sets and Generalized Boolean Retrieval Systems. *International Journal of Man-Machine Studies*, 19(1):45–56.

Krovetz, R. (1993). Viewing Morphology as an Inference Process. *In Proceedings of the 16th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Pittsburgh, États-Unis.

Larousse (1998). *Le Petit Larousse 1999.* Larousse.

Lee, J. H. (1995). Combining Multiple Evidence from Different Properties of Weighting Schemes. *In Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Seattle, États-Unis.

Lee, J. H. (1997). Analyses of Multiple Evidence Combination. *In Proceedings of the 20th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Philadelphie, États-Unis.

Lennon, M., Pierce, D. S., Tarry, B. D. et Willett, P. (1981). An Evaluation of some Conflation Algorithms for Information Retrieval. *Journal of Information Science*, 3(1):177–183.

Lepage, Y. (2003). *De l'analogie, rendant compte de la communication en linguistique.* Mémoire d'habilitation à diriger des recherches, Université de Grenoble 1, Grenoble, France.

Lewis, D. D. (1992). An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *In Proceedings of the 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Copenhague, Danemark.

Lewis, D. D. et Croft, W. B. (1990). Term Clustering of Syntactic Phrases. *In Proceedings of the 13th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Bruxelles, Belgique.

Loupy, C. d. (2000). *Évaluation de l'apport de connaissances linguistiques en désambiguïsation sémantique et recherche documentaire.* Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, Avignon, France.

Loupy, C. d. et Bellot, P. (2000). Evaluation of Document Retrieval System and Query Difficulty. *In Proceedings of Using Evaluation within HLT Programs : Results and Trends*, Athènes, Grèce.

Lovins, J. B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11(1):22–31.

Luhn, H. (1978). The Automatic Creation of Literature Abstract. *IBM Journal of Research and Development*, 2(2):159–165.

Lundquist, C., Grossman, D. A. et Frieder, O. (1997). Improving Relevance Feedback in the Vector Space Model. *In Proceedings of the 6th ACM Conference on Information and Knowledge Management (CIKM)*, Las Vegas, États-Unis.

Macdonald, C., He, B. et Ounis, I. (2005). Predicting Query Performance in Intranet Search. *In Proceedings of Predicting Query Difficulty - Methods and Applications Workshop, ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador de Bahia, Brésil.

Maisonnasse, L. (2005). Vers l'exploitation d'analyse de dépendances en recherche d'information précise. *In Proceedings of 23ème conférence en informatique des organisations et systèmes d'information et de décision (INFORSID)*, Grenoble, France.

Mandala, R., Tokunaga, T., Tanaka, H., Akitoshi, O. et Satoh, K. (1998). Ad Hoc Retrieval Experiments using WORDNET and Automatically Constructed Thesauri. *In Proceedings of the 7th International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

MANNING, D. et SCHÜTZE, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press.

MATSUMURA, A., TAKASU, A. et ADACHI, J. (2000). The Effect of Information Retrieval Method using Dependency Relationship between Words. *In Proceedings of 6ème conférence internationale sur la recherche d'informations assistée par ordinateur (RIAO)*, Paris, France.

MCCABE, M. C., CHOWDHURY, A., GROSSMAN, D. A. et FRIEDER, O. (1999). A Unified Environment for Fusion of Information Retrieval Approaches. *In Proceedings of the 8th ACM Conference on Information and Knowledge Management (CIKM)*, Kansas City, États-Unis.

METZLER, D. P. et HASS, S. W. (1989). The Constituent Object Parser : Syntactic Structure Matching for Information Retrieval. *ACM Transactions on Information Systems*, 7(3):296–316.

MIHALCEAN, R. et MOLDOVAN, D. I. (2000). Semantic Indexing using WORDNET Senses. *In Proceedings of Workshop on IR and NLP, 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, Hong-Kong, Chine.

MILLER, D., LEEK, T. et SCHWARTZ, R. (1999). A Hidden Markov Model Information Retrieval System. *In Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Berkeley, États-Unis.

MITRA, M., BUCKLEY, C., SINGHAL, A. et CARDIE, C. (1997). An Analysis of Statistical and Syntactic Phrases. *In Proceedings of 5ème conférence internationale sur la recherche d'informations assistée par ordinateur (RIAO)*, Montréal, Canada.

MIZARRO, S. (1997). Relevance, the whole History. *Journal of the American Society for Information Science*, 48(9):810–832.

MOREAU, F. et SÉBILLOT, P. (2005). Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport de recherche, IRISA, Rennes, France.

MORIN, É. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat, Université de Nantes, Nantes, France.

MOTHE, J. et TANGUY, L. (2005). Linguistic Features to Predict Query Difficulty - a Case Study on previous TREC Campaigns. *In Proceedings of Predicting query difficulty - methods and applications Workshop, ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador de Bahia, Brésil.

MOULINIER, I., MCCULLOH, J. A. et LUND, E. (2000). West Group at CLEF 2000 : Non-English Monolingual Retrieval. *Lecture Notes in Computer Science*, 2069:176–187.

Mounir, S. L., Goharian, N., Mahoney, M., Salem, A. et Frieder, O. (1998). Fusion of Information Retrieval Engines (fire). *In Proceedings of the International Conference on Parallel and Distributed Processing Technologies and Applications (PDPTA)*, Las Vegas, États-Unis.

Müller, H. M., Kenny, E. E. et Sternberg, P. W. (2004). textpresso : An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biology*, 2(11):e309.

Namer, F. (2000). flemm : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues*, 41(2):523–547.

Nie, J.-Y. (1990). *Un modèle logique général pour les systèmes de recherche d'informations. Application au prototype RIME.* Thèse de doctorat, Université Joseph Fourier, Grenoble, France.

Nie, J.-Y. (Cours en ligne version 2006). Indexation de documents. http ://www.iro.umontreal.ca/%7Enie/IFT6255/Indexation.html.

Oard, D. W., Levow, G. A. et Cabezas, C. I. (2001). CLEF Experiments at Maryland : Statistical Stemming and Backoff Translation. *In Proceedings of Workshop of Cross-Language Evaluation Forum (CLEF)*, Lisbonne, Portugal.

Ogawa, Y., Morita, T. et Kobayashi, K. (1991). A Fuzzy Document Retrieval System using the Keyword Connection Matrix and a Learning Method. *Fuzzy Sets and Systems*, 39:163–179.

Paice, C. D. (1990). Another Stemmer. *SIGIR Forum*, 24(3):56–61.

Parker, D. B. (1985). Learning Logic. Rapport technique, Center for Computational Research in Economics and Management Science, MIT, Cambridge, États-Unis.

Peat, H. J. et Willett, P. (1991). The Limitations of Term Cooccurrence Data for Query Expansion in Document Retrieval Systems. *Journal of the American Society for Information Science*, 42(5):378–383.

Pedersen, T. et Bruce, R. (1997). A New Supervised Learning Algorithm for Word Sense Disambiguation. *In Proceedings of the 14th National Conference on Artificial Intelligence, AAAI*, Providence, États-Unis.

Pedersen, T., Patwardhan, S. et Michelizzi, J. (2004). wordnet : :similarity - Measuring the Relatedness of Concepts. *In Proceedings of 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Boston, États-Unis.

Perez-Carballo, J. et Strzalkowski, T. (2000). Natural Language Information Retrieval : Progress Report. *Information Processing and Management*, 36(1):155–178.

Pichon, R. et Sébillot, P. (2000). From Corpus to Lexicon : From Contexts to Semantic Features. *In* Lewandowska-Tomaszczyk, B. et Melia, P., éditeurs : *PALC'99 : Practical Applications in Language Corpora*, pages 375–389. Peter Lang.

Piwowarski, B. (2003). *Techniques d'apprentissage pour le traitement d'informations structurées : application à la recherche d'information.* Thèse de doctorat, Université Pierre et Marie Curie, Paris, France.

Polguère, A. (2003). *Lexicologie et sémantique lexicale - Notions fondamentales.* Presse de l'Université de Montréal.

Ponte, J. et Croft, W. B. (1998). A Language Modeling Approach to Information Retrieval. *In Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Melbourne, Australie.

Popovic, M. et Willett, P. (1992). The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. *Journal of the American Society for Information Science*, 43(5):384–390.

Porter, M. F. (1980). An Algorithm for Suffix Stripping. *Program*, 14(1):130–137.

Qiu, Y. et Frei, H. P. (1995). Improving the Retrieval Effectiveness by a Similarity Thesaurus. Rapport interne, Department of Computer Science, ETH Zürich, Zürich, Suisse.

Raghavan, R. R. et Wong, S. K. M. (1986). A Critical Analysis of Vector Space Model for Information Retrieval. *Journal of the American Society for Information Science*, 37(2):279–287.

Rajashekar, T. B. et Croft, W. B. (1995). Combining Automatic and Manual Index Representations in Probabilistic Retrieval. *Journal of American Society of Information Science*, 46(4):272–283.

Rajman, M., Besançon, R. et Chappelier, J.-C. (2000). Le modèle dsir : une approche à base de sémantique distributionnelle pour la recherche documentaire. *Traitement automatique des langues*, 41(2):549–578.

Ramshaw, L. et Marcus, M. (1995). Text Chunking using Transformation-Based Learning. *In Proceedings of the 3th Workshop on Very Large Corpora*, Somerset, États-Unis.

Renda, E. M. et Straccia, U. (2003). Web Metasearch : Rank *vs* Score Based Rank Aggregation Methods. *In Proceedings of the 18th Annual ACM Symposium on Applied Computing*, Melbourne, États-Unis.

Ribeiro, B. A. N. et Muntz, R. (1996). A Belief Network Model for IR. *In Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Zürich, Suisse.

RICHARDSON, R., SMEATON, A. F. et MURPHY, J. (1996). Using WORDNET for Conceptual Distance Measurement. *In Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group*, Londres, Angleterre.

RIEGEL, M., PELLAT, J. C. et RIOUL, R. (1999). *Grammaire méthodique du français*. PUF, collection linguistique nouvelle.

RIJSBERGEN, C. J. v. (1977). A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. *Journal of Documentation*, 33(2):106–119.

RIJSBERGEN, C. J. v. (1979). *Information Retrieval*. Butterworths.

RIJSBERGEN, C. J. v. (1986). A New Theoretical Framework for Information Retrieval. *In Proceedings of the 9th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Pise, Italie.

ROBERTSON, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304.

ROBERTSON, S. E., RIJSBERGEN, C. J. v. et PORTER, M. F. (1981). Probabilistic Models of Indexing and Searching. *In Proceedings of the 3rd ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Cambridge, Angleterre.

ROBERTSON, S. E. et SPÄRCK JONES, K. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146.

ROBERTSON, S. E. et WALKER, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *In Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Dublin, Irlande.

ROBERTSON, S. E., WALKER, S. et HANCOCK-BEAULIEU, M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. *In Proceedings of the 7th International Conference on Text Retrieval (TREC)*.

ROCCHIO, J. J. (1971). Relevance Feedback in Information Retrieval. *In* SALTON, G., éditeur : *The* SMART *Retrieval System : Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall.

RUMELHART, D. E., HINTON, G. E. et WILLIAMS, R. J. (1986). Learning Internal Representations by Error Propagation. *In* RUMELHART, D. et MCCLELLAND, J., éditeurs : *Parallel Distributed Processing*. MIT Press.

RUTHVEN, I. et LALMAS, M. (2003). A Survey on the Use of Relevance Feedback for Information Access Systems. *Knowledge Engineering Review*, 18(1):95–145.

SALTON, G. (1971). *The* SMART *Retrieval System : Experiments in Automatic Document Processing*. Prentice-Hall.

SALTON, G. (1975). *A Theory of Indexing*. Society for Industrial and Applied Mathematics.

SALTON, G. (1992). The State of Retrieval System Evaluation. *Information Processing and Management*, 28(4):441–449.

SALTON, G. et BUCKLEY, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523.

SALTON, G. et BUCKLEY, C. (1990). Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science*, 41(4):288–297.

SALTON, G., FOX, E. et WU, H. (1983). Extended Boolean Information Retrieval. *Communications of the ACM*, 31(2):1002–1036.

SALTON, G., WONG, A. et YANG, C. (1975). A Vector Space Model for Automatic. *Communications of the ACM*, 18(11):613–620.

SANDERSON, M. (1997). *Word Sense Disambiguation and Information Retrieval*. Thèse de doctorat, Université de Glasgow, Glasgow, Écosse.

SARACEVIC, T. (1996). Relevance Reconsidered. *In Proceedings of the 2nd International Conference on Conceptions of Library and Information Science (COLIS)*, Copenhague, Danemark.

SAUVAGNAT, K. (2005). *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

SAVOY, J. (1993). Stemming of French Words Based on Grammatical Categories. *Journal of the American Society for Information Science*, 44(1):1–9.

SAVOY, J. (1999). A Stemming Procedure and Stopword List for General French Corpora. *Journal of the American Society for Information Science*, 50(10):944–952.

SAVOY, J. (2002). Morphologie et recherche d'information. Rapport technique, Institut interfacultaire d'informatique, Université de Neuchâtel, Neuchâtel, Suisse.

SAVOY, J., CALVÉ, A. L. et VRAJITORU, D. (1997). Report on the TREC-5 Experiment : Data Fusion and Collection Fusion. *In Proceedings of the 5th International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

SCHANK, R. C. (1972). Dependency : A Theory of Natural Language Understanding. *Cognitive Psychology*, 3(4):532–631.

SCHMID, H. (1997). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In* JONES, D. et SOMERS, H., éditeurs : *New Methods in Language Processing*, pages 154–164. UCL Press.

SCHÜTZE, H. et PEDERSEN, J. O. (1995). Information Retrieval Based on Word Senses. *In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, États-Unis.

SINGHAL, A. K. (1997). *Term Weighting Revisited.* Thèse de doctorat, Université de Cornell, New-York, États-Unis.

SMEATON, A. F. (1999). Using NLP or NLP Ressources for Information Retrieval Tasks. *In* STRZALKOWSKI, T., éditeur : *Natural Language Information Retrieval*, pages 99–111. Kluwer Academic Publishers.

SMEATON, A. F., KELLEDY, F. et O'DONELL, R. (1995). TREC-4 Experiments at Dublin City University : Thresholding Posting Lists, Query Expansion with WORDNET and POS Tagging of Spanish. *In Proceedings of the 4th International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

SONG, F. et CROFT, W. B. (1999). A General Language Model for Information Retrieval. *In Proceedings of the 8th ACM Conference on Information and Knowledge Management (CIKM)*, Kansas City, États-Unis.

SOWA, J. F. (1984). *Conceptual Structures : Information Processing in Mind and Machine.* Addison-Wesley.

SPÄRCK JONES, K. (1999). What is the Role of NLP in Text Retrieval ? *In* STRZALKOWSKI, T., éditeur : *Natural Language Information Retrieval*, pages 1–24. Kluwer Academic Publishers.

SPÄRCK JONES, K. (2000). Summary Performance Comparisons, TREC-2 through TREC-8. *In* VOORHEES, E. M. et HARMAN, D., éditeurs : *The 8th Text Retrieval Conference (TREC-8)*, pages B–1. NIST Special Publication.

SPÄRCK JONES, K. et RIJSBERGEN, C. J. v. (1975). Report on the Need for and Provision of an Ideal Information Retrieval Test Collection. Rapport technique, Computer Laboratory, Université de Cambridge, Cambridge, États-Unis.

SPÄRCK JONES, K., WALKER, S. et ROBERTSON, S. E. (2000). A Probabilistic Model for Information Retrieval : Development and Comparative Experiments, part 1 and 2. *Information Processing and Management*, 36(6):779–840.

SRIKANTH, M. et SRIHARI, R. (2002). Biterm Language Models for Document Retrieval. *In Proceedings of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, New-York, États-Unis.

STRZALKOWSKI, T., LIN, F., WANG, J. et PEREZ-CARBALLO, J. (1999). Evaluating Natural Language Processing Techniques in Information Retrieval. *In* STRZALKOWSKI, T., éditeur : *Natural Language Information Retrieval*, pages 113–145. Kluwer Academic Publishers.

TOWELL, G. G. et VOORHEES, E. M. (1998). Disambiguating Highly Ambiguous Words. *Computational Linguistics*, 24(1):125–145.

TURTLE, H. T. et CROFT, W. B. (1991). Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions On Information Systems*, 9(3):187–222.

UZUNER, O., KATZ, B. et YURET, D. (1999). Word Sense Disambiguation for Information Retrieval. *In Proceedings of the 16th National Conference on Artificial Intelligence (AAAI)*, Orlando, États-Unis.

VILARES-FERRO, J., BARCALA, F. M. et ALONSO, M. A. (2002). Using Syntactic Dependency-Pairs Conflation to Improve Retrieval Performance in Spanish. *In Proceedings of the 3th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Mexico, Mexique.

VOGT, C. C. et COTTRELL, G. W. (1999). Fusion via a Linear Combination of Scores. *Information Retrieval*, 1(3):151–173.

VOORHEES, E. M. (1998). Using WORDNET for Text Retrieval. *In* FELLBAUM, C., éditeur : WORDNET *: An Electronic Lexical Database*, pages 285–303. The MIT Press.

VOORHEES, E. M. (1999). Natural Language Processing and Information Retrieval. *In* PAZIENZA, M. T., éditeur : *Information Extraction : Towards Scalable, Adaptable Systems*, pages 32–48. Springer.

VOORHEES, E. M. et HARMAN, D. (1996). Overview of the 5th Text Retrieval Conference. http ://trec.nist.gov/pubs/trec5/papers/overview.ps.gz.

WILKINSON, R., HINGSTON, P. et OSBORN, T. (1992). Incorporating the Vector Space Model in a Neural Network used for Document Retrieval. *Library Hi Tech*, 10(12):69–75.

WONG, S. K. M., ZIARKO, W. et WONG, C. N. (1985). Generalized Vector Space Model in Information Retrieval. *In Proceedings of the 9th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Pise, Italie.

WOODS, W. A. et AMBROZIAK, J. (1998). Natural Language Technology in Precision Content Retrieval. *In Proceedings of the International Conference on Natural Language Processing and Industrial Applications, (NLP+IA)*, Moncton, Canada.

XU, J. et CROFT, W. B. (1998). Corpus-Based Stemming Using Cooccurrence of Word Variants. *ACM Transactions on Information Systems*, 16(1):61–81.

Xu, J. et Croft, W. B. (2000). Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112.

Yom-Tom, E., Fine, S., Carmel, D., Darlow, A. et Amitay, E. (2005). Learning to Estimate Query Difficulty : Including Applications to Missing Content Detection and Distributed Information Retrieval. *In Proceedings of the 28th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador de Bahia, Brésil.

Zhai, C., Tong, X., Milic-Frayling, N. et Evans, D. A. (1997). Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report. *In Proceedings of the 5th International Conference on Text Retrieval (TREC)*, Gaithersburg, États-Unis.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least*. Addison-Wesley Press.

Zobel, J. (1998). How Reliable are the Results of Large-Scale Iinformation Retrieval Experiments ? *In Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Melbourne, Australie.

Zweigenbaum, P., Grabar, N. et Darmoni, S. (2001). Apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée. *In Proceedings of 8ème conférence annuelle sur le traitement automatique des langues naturelles (TALN)*, Tours, France.

Zweigenbaum, P. et Menelas, C. (1994). menelas : An Access System for Medical Records Using Natural Language. *Computer Methods and Programs in Biomedicine*, 45(1):117–120.

Table des figures

Résumé

La principale difficulté des systèmes de recherche d'information (SRI) est d'établir une correspondance entre l'information recherchée par un utilisateur et celle contenue dans leur base documentaire. Pour y parvenir, ils tentent généralement un appariement des mots de la requête posée avec ceux représentant le contenu des documents. Un tel mécanisme, fondé sur une simple comparaison de chaînes de caractères, ne permet cependant pas de prendre en compte le fait qu'un même mot peut posséder plusieurs sens et qu'une même idée peut être formulée de différentes manières. Pour pallier ces difficultés, une solution assez naturelle est de se tourner vers le traitement automatique des langues (TAL) qui, en considérant les mots non comme des chaînes de caractères mais comme des entités linguistiques à part entière, doit offrir un appariement requête-document plus pertinent. Les résultats des nombreux travaux proposant d'enrichir la RI par des informations linguistiques sont toutefois souvent décevants, peu tranchés et contradictoires. Pour comprendre ces faibles résultats et savoir comment les améliorer, nous abordons le couplage TAL-RI sous des angles nouveaux. Contrairement aux autres études, nous choisissons d'exploiter pleinement la richesse de la langue en combinant plusieurs informations linguistiques appartenant aux niveaux morphologique, syntaxique et sémantique. Afin de tester l'intérêt de coupler ces informations, nous proposons une plate-forme intégrant en parallèle ces multiples indices ; elle conduit à montrer l'apport significatif et tranché de plusieurs de ces connaissances, et, via une analyse originale des corrélations qu'elles présentent, des cas de complémentarité intéressants. Grâce à une méthode d'apprentissage supervisé qui fusionne les listes de résultats fournis par chaque index linguistique et s'adapte automatiquement aux caractéristiques des requêtes, nous prouvons, par des résultats plus stables qu'habituellement, le gain effectif du couplage d'informations linguistiques multi-niveaux. Enfin, nous proposons une méthode novatrice d'acquisition par apprentissage non supervisé d'informations morphologiques qui permet d'accroître encore l'impact de ces connaissances efficaces sur les performances de notre SRI. Nous montrons ainsi qu'en construisant des outils plus souples et plus adaptés aux contraintes de la RI, l'apport du TAL dans ce domaine est réel.

Abstract

Information retrieval systems (IRSs) aim at establishing a relationship between users' information needs and the information contained in documents. To this end, a commonly used method consists of making a simple match between query terms and document words. IRSs face two problems with such a mechanism. The first problem is related to polysemy : a single term may have different meanings and represent various concepts. The second and dual issue reflects the fact that a single idea may be expressed in different forms. To overcome these limitations, a more natural solution is to perform a linguistic analysis of both documents and queries, using natural language processing (NLP) techniques. This allows one to consider each word as a single linguistic entity rather than as a simple string of characters, thus providing a more relevant document-query match. However, many previous studies that have tried to enrich IRSs with linguistic information have often resulted in disappointing unclear and and contradictory outputs. In order to better understand and improve upon these weak results, we propose a new approach for coupling NLP-IR. In contrast with other studies, we choose to fully exploit the richness of language by combining several levels of linguistic information : morphological, syntactic and semantic. To test the proposition of linking these various knowledges, we have designed a test platform which integrates them in parallel within the same IRSs ; this serves to demonstrate the clear and significant contribution of several types of information (especially morphological and semantic) and, via an original analysis of the correlations between the various linguistic index, it has highlighted some interesting cases of a complementary nature. Through a supervised machine-learning technique that merges the list of documents produced with each linguistic index, and automatically adapts its behavior to the query's characteristics, we prove how combining multilevel linguistic information can provide better overall results that are also far more stable than comparable tests. Finally, we propose a new method for the acquisition of morphological variants based on unsupervised learning techniques, which provides an even greater impact of this efficient knowledge on the performance of our IRS system. We show that by introducing more flexible tools that are better adapted to the constraints of IR, NLP can make a real contribution to this area.